第十章　数据拟合与插值

§10.1 引言

在解决实际问题的生产（或工程）实践和科学实验过程中，通常需要通过研究某些变量之间的函数关系来帮助我们认识事物的内在规律和本质属性，而这些变量之间的未知函数关系又常常隐含在从试验、观测得到的一组数据之中。因此，能否根据一组试验观测数据找到变量之间相对准确的函数关系就成为解决实际问题的关键。

例如在工程实践和科学实验中，常常需要从一组试验观测数据(x_i, y_j)，$i = 0,1,\cdots,n$之中找到自变量x与因变量y之间的函数关系，一般可用一个近似函数$y = f(x)$来表示。函数$y = f(x)$的产生办法因观测数据和要求不同而异，通常可采用数据拟合与函数插值两种办法来实现。

数据拟合主要是考虑到观测数据受随机观测误差的影响，进而寻求整体误差最小、能较好反映观测数据的近似函数$y = f(x)$，此时并不要求所得到的近似函数$y = f(x)$满足$y_i = f(x_i)$，$i = 0,1,\cdots,n$。

函数插值则要求近似函数$y = f(x)$在每一个观测点x_i处一定要满足$y_i = f(x_i)$，$i = 0,1,\cdots,n$。在这种情况下，通常要求观测数据相对比较准确，即不考虑观测误差的影响。

在实际问题中，通过观测数据能否正确揭示某些变量之间的关系，进而正确认识事物的内在规律与本质属性，往往取决于两方面因素。其一是观测数据的准确性或准确程度，这是因为在获取观测数据的过程中一般存在随机测量误差，导致所讨论的变量成为随机变量。其二是对观测数据处理方法的选择，即到底是采用插值方法还是用拟合方法，插值方法之中、拟合方法之中又选用哪一种插值或拟合技巧来处理观测数据。插值问题忽略了观测误差的影响，而拟合问题则考虑了观测误差的影响。但由于观测数据客观上总是存在观测误差，而拟合函数大多数情况下是通过经验公式获得的，因此要正确揭示事物的内在规律，往往需要对大量的观测数据进行分析，尤为重要的是进行统计分析。统计分析的方法有许多，如方差分析、回归分析等。数据拟合虽然较有效地克服了随机观测误差的影响，但从数理统计的角度看，根据一个样本计算出来的拟合函数（系数），只是拟合问题的一个点估计，还不能完全说明其整体性质。因此，还应该对拟合函数作区间估计或假设检验，如果置信区间太大或包含零点，则由计算得到的拟合函数系数的估计值就毫无意义。这里所采用的统计分析方法就是所谓的回归分析。另外还可用方差分析的方法对模型的误差作定量分析。

对于插值方法，本章在第二节中简单介绍最常用的插值法的基本结论及其Matlab实现问题。由于数据拟合问题必须作区间估计或假设检验，所以除了在本章第三节、第四节中介绍最基本的数据拟合方法——最小二乘法的基本结论及其Matlab实现问题外，我们在第五节中专门介绍了对数值拟合问题进行区间估计或假设检验的统计方法，即介绍回归分析方法及其Matlab实现。

数据处理问题通常情况下只是某个复杂实际问题的一个方面或部分内容，因而这里所介绍

的数据处理方法——函数插值和数据拟合的方法（包括回归分析）通常只能解决实际问题中的部分问题——计算问题。一般来说，对实际问题进行数学建模需要用到多方面知识，只有很少的情况下可以单独使用本章所介绍的内容，故我们只在本章最后一节以修改后的美国91年数学建模A题为例说明如何使用数值计算知识建立数学模型，从而解决实际问题的方法。

§10.2 插值方法

在实际问题中所遇到的插值问题一般分为一维插值问题和二维插值问题。本节主要介绍最常用的一维插值方法及其一些简单结果。

一维插值问题的数学描述为：已知某一函数 $y=g(x)$（$g(x)$ 的解析表达式可能十分复杂，也可以是未知的）在区间 $[a,b]$ 上 $n+1$ 个互异点 x_j 处的函数值 y_j，$j=0,1,\cdots,n$，还知道 $g(x)$ 在 $[a,b]$ 上有若干阶导数，如何求出 $g(x)$ 在 $[a,b]$ 上任一点 x 的近似值。

一维插值方法的基本思想是：根据 $g(x)$ 在区间 $[a,b]$ 上 $n+1$ 个互异点 x_j（称为节点）的函数值 y_j，$j=0,1,\cdots,n$，求一个足够光滑、简单便于计算的函数 $f(x)$（称为插值函数）作为 $g(x)$ 的近似表达式，使得

$$f(x_j)=y_j，\quad j=0，1，\cdots，n。\tag{10.1}$$

然后计算 $f(x)$ 在区间 $[a,b]$(称为插值区间)上点 x（称为插值点）的值作为原函数 $g(x)$（称为被插函数）在此点的近似值。求插值函数 $f(x)$ 的方法称为插值方法，（10.1）称为插值条件。

代数多项式比较简单，常用多项式作为插值函数。

一　插值多项式的存在唯一

假设 $f(x)$ 是一个满足插值条件（10.1）的次数不超过 n 的代数多项式，即

$$f(x)=a_0+a_1x+\cdots+a_nx^n \tag{10.2}$$

为满足（10.1）的插值函数，则 $f(x)$ 的 $n+1$ 个待定系数 a_0，a_1，…，a_n 满足

$$\begin{cases} a_0+a_1x_0+a_2x_0^2+\cdots+a_nx_0^n=y_0, \\ a_0+a_1x_1+a_2x_1^2+\cdots+a_nx_1^n=y_1, \\ \qquad\cdots\qquad\cdots \\ a_0+a_1x_n+a_2x_n^2+\cdots+a_nx_n^n=y_n. \end{cases} \tag{10.3}$$

记此方程组的系数矩阵为 A，则

$$\det(A)=\begin{vmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{vmatrix}$$

是范德蒙(Vandermonde)行列式。当 $x_0,x_1,\cdots,x_n$ 互不相同时，此行列式值不为零。因此方程组（10.3）有唯一解。这表明，只要 $n+1$ 个插值节点 $x_0,x_1,\cdots,x_n$ 互异，满足插值条件（10.1）的

插值多项式（10.2）存在唯一。

从几何上看，上述多项式插值就是过$n+1$个数据点(x_j, y_j)，j=0，1，…，n作一条多项式曲线$y=f(x)$近似原有曲线$y=g(x)$。

当$x\in[a,b]$且$x\neq x_j\ (j=0,1,\cdots,n)$时，$g(x)\approx f(x)$，称被插函数$g(x)$与插值函数（多项式）$f(x)$之间的差

$$R_n(x)=g(x)-f(x)$$

为插值函数（多项式）$f(x)$的截断误差，或插值余项。显然，插值多项式的存在唯一与节点$x_0,x_1,\cdots,x_n$的次序无关。

二　拉格朗日插值法

1 拉格朗日插值的基本结果

用多项式函数（10.2）作为插值函数时，希望通过解方程组（10.3）而得到待定系数a_0，a_1，…，a_n的做法当n比较大时是不现实的。方便的方法是先构造一组基函数

$$l_i(x)=\frac{(x-x_0)\cdots(x-x_{i-1})(x-x_{i+1})\cdots(x-x_n)}{(x_i-x_0)\cdots(x_i-x_{i-1})(x_i-x_{i+1})\cdots(x_i-x_n)}$$，i=0，1，…，n。

显然$l_i(x)$是n次多项式且满足

$$l_i(x_j)=\begin{cases}0, & j\neq i,\\ 1, & j=i,\end{cases}\qquad i,j=0,1,\cdots,n$$。

称n次多项式$l_i(x)$为节点$x_0,x_1,\cdots,x_n$上的n次插值基函数。令

$$L_n(x)=\sum_{i=0}^{n} y_i l_i(x)$$，

则多项式$L_n(x)$显然满足插值条件$L_n(x_i)=y_i\quad(i=0,1,\cdots,n)$，从而$L_n(x)$为插值多项式。插值多项式$L_n(x)$又称为$n$次拉格朗日（Lagrange）插值多项式，由方程组（10.3）解的存在唯一性，$n+1$个互异节点的n次拉格朗日插值多项式$L_n(x)$存在唯一。

当$g(x)$在$[a,b]$上充分光滑时，利用罗尔（Rolle）定理可推出：对于任意$x\in[a,b]$，插值多项式$L_n(x)$的余项

$$R_n(x)=g(x)-L_n(x)=\frac{g^{(n+1)}(\xi)}{(n+1)!}\omega_{n+1}(x),\xi\in(a,b)$$，

其中$\omega_{n+1}(x)=\prod_{j=0}^{n}(x-x_j)$。若可以估计

$$|g^{(n+1)}(\xi)|\leq M_{n+1}$$，

则可以得到n次拉格朗日插值的余项估计

$$|R_n(x)|\leq\frac{M_{n+1}}{(n+1)!}|\omega_{n+1}(x)|$$。

实际上，因为M_{n+1}常难以确定，所以上式并不能给出精确的误差估计。

2 拉格朗日插值的Matlab实践

Matlab中没有现成的拉格朗日插值函数，必须编写一个M文件实现拉格朗日插值。

设 n 个节点数据以数组 $x0, y0$ 输入（注意Matlab的数组下标从1开始），m 个插值点以数组 x 输入，输出数组 y 为 m 个插值。编写一个名为lagrange.m的M文件：

```
function y=lagrange(x0,y0,x);
n=length(x0);m=length(x);
for i=1:m
   z=x(i);
   s=0.0;
   for k=1:n
      p=1.0;
      for j=1:n
         if j~=k
            p=p*(z-x0(j))/(x0(k)-x0(j));
         end
      end
      s=p*y0(k)+s;
   end
   y(i)=s;
end
```

三　分段线性插值法

1 高次插值多项式的龙格（Runge）现象

用拉格朗日插值多项式 $L_n(x)$ 作为区间 $[a,b]$ 上连续函数 $g(x)$ 的近似函数，在大多数情况下，$L_n(x)$ 的次数越高，逼近 $g(x)$ 的效果就越好，即误差 $|R_n(x)|$ 越小。但是对于高次多项式插值问题而言，往往会造成插值多项式 $L_n(x)$ 的收敛性与稳定性变差，反而逼近效果不理想，甚至发生龙格现象，即在理论上并不能保证当 n 趋于无穷大时 $L_n(x)$ 在 $[a,b]$ 上处处收敛于 $g(x)$。关于这一点，在20世纪初就为龙格（Runge）所发现：在 $[-1,1]$ 上用 $n+1$ 个等距节点作被插值函数 $y(x)=1/(1+25x^2)$ 的插值多项式 $L_n(x)$，则随着 n 的增大，$L_n(x)$ 振荡越来越大。计算结果与理论证明表明，当 n 趋于无穷大时，$L_n(x)$ 在区间中部收敛于 $y(x)$，但是对于满足条件 $0.726\cdots \le |x| < 1$ 的 x，$L_n(x)$ 并不收敛于 $y(x)$。

上述例子说明，使用高次多项式插值是危险的，因此在实际计算中不能使用高次插值。这反而启发我们使用分段插值方法，即将区间 $[a,b]$ 分成一些小区间，在每一个小区间上用低次多项式进行插值，在整个插值区间 $[a,b]$ 上就得到一个分段低次多项式插值函数。区间的剖份可以是任意的，各小区间上插值多项式的次数的选取也可按具体问题的要求而选择。分段低次多项式插值通常有较好的收敛性和稳定性，算法简单，但插值函数光滑性变差。常用的分段多项式插值法有两类：一类是下面将要介绍的分段线性插值法；另一类是后面将要介绍的三次样条插值法。

2 分段线性插值法

假设区间$[a,b]$上的连续函数$g(x)$在$n+1$个节点$a=x_0<x_1<\cdots<x_n=b$上的函数值$g(x_j)=y_j$，j=0，1，…，n。则得到xy平面上的$n+1$个数据点(x_j,y_j)。连接相邻数据点(x_{j-1},y_{j-1})、(x_j,y_j)得到n条线段，它们组成一条折线。把区间$[a,b]$上这n条折线段表示的函数称为被插函数$g(x)$关于这$n+1$个数据点的分段线性插值函数，记作$I(x)$，则$I(x)$具有如下性质：

(1) $I(x)$可以分段表示，在每个小区间$[x_{j-1},x_j]$上，它是线性函数，即

$$I(x)=y_{j-1}\frac{x-x_j}{x_{j-1}-x_j}+y_j\frac{x-x_{j-1}}{x_j-x_{j-1}},\quad x_{j-1}\le x\le x_j;$$

(2) $I(x_j)=y_j$，j=0，1，…，n；

(3) $I(x)$在$[a,b]$上连续。

若构造插值基函数

$$l_i(x)=\begin{cases}\dfrac{x-x_{i-1}}{x_i-x_{i-1}}, & x\in[x_{i-1},x_i]\ (i=0\text{时舍去}),\\ \dfrac{x-x_{i+1}}{x_i-x_{i+1}}, & x\in[x_i,x_{i+1}]\ (i=n\text{时舍去}),\\ 0, & \text{其它},\end{cases}$$

则$I(x)=\sum_{j=0}^{n}y_jl_j(x)$。

当$g(x)$在$[a,b]$上连续时，分段线性插值函数$I(x)$具有良好的收敛性，即

$$\lim_{h\to 0}I(x)=g(x),\quad x\in[a,b],$$

而且当$g(x)$在$[a,b]$上二阶导数连续时，对于任意$x\in[a,b]$有

$$|R(x)|=|g(x)-I(x)|\le\frac{M_2}{8}h^2,$$

其中$h=\max_{1\le j\le n}\{x_j-x_{j-1}\}$，$M_2=\max_{a\le x\le b}\{|g''(x)|\}$。

用$I(x)$计算x点的插值时，只用到x左右的两个节点，计算量与节点个数$n+1$无关。但n越大，分段越多，插值误差越小。实际上用数据点作插值计算时，分段线性插值就足够了，如数学、物理中用的特殊函数表，数理统计中用的概率分布表等。

3 分段线性插值的Matlab实现

用Matlab实现分段线性插值不需要编制函数程序，Matlab中有现成的一维插值函数interp1。

```
y=interp1(x0,y0,x,'method')
```

method指定插值的方法，默认为线性插值。其值可为：

'nearest' 最近项插值

'linear' 线性插值

'spline' 立方样条插值

'cubic'　立方插值。

所有的插值方法要求x0是单调的。

当x0为等距时可以用快速插值法，使用快速插值法的格式为 '*nearest'、'*linear' 、'*spline' 、'*cubic'。

四　三次样条插值法

1 三次样条的基本结果

分段线性插值函数在节点处的一阶导数一般不存在，光滑性不高，这影响了它在诸如机械加工等领域（希望插值曲线光滑）中的应用。例如许多工程技术中提出的计算问题对插值函数的光滑性有较高要求，如飞机的机翼外形，内燃机的进、排气门的凸轮曲线，都要求曲线具有较高的光滑程度，不仅要连续，而且要有连续的曲率。在船舶、飞机等设计中，面对已知的一些数据点 (x_0,y_0)，(x_1,y_1)，…，(x_n,y_n)，绘图员的做法是：首先将这些数据点描绘在平面图纸上，再把一根富有弹性的细直条（称为样条）弯曲，使其一边通过这些数据点，用压铁固定细直条的形状，沿样条边沿绘出一条光滑曲线。往往要用几根样条分段完成上述工作，这时应当让样条在连接处也保持光滑。

根据力学理论进行分析，这样画出的曲线，在相邻两数据点之间实际上是次数不高于3的多项式，在整个区间上有连续的曲率。对绘图员用样条画出的曲线进行数学模拟，导出了样条函数的概念。

定义于区间 $[a,b]$ 上的分段函数 $S(x)$，若满足

(1) $S(x)$ 在每一小区间 $[x_{j-1},x_j]$ 上是一个三次多项式函数；

(2) $S(x)$ 在整个区间 $[a,b]$ 上有连续的二阶导数。

则称 $S(x)$ 为 $[a,b]$ 上关于剖分 $a=x_0<x_1<\cdots<x_n=b$ 的一个三次样条函数。

从而三次样条插值问题为：给定函数 $g(x)$ 在 $n+1$ 个节点 x_0，x_1，…，x_n 得函数值 y_0，y_1，…，y_n，求一个三次样条函数 $S(x)$，使其满足

$$S(x_j)=y_j,\quad j=0，1，\cdots，n,$$

其中函数 $S(x)$ 称为 $g(x)$ 的三次样条插值函数。

对于三次样条插值函数而言，在每一个小区间上有四个待定参数，共有参数 $4n$ 个。而在每一小区间上，由插值条件可得到2个方程，共计 $2n$ 个方程。光滑性要求 $S(x)$ 在每一内部节点一阶、二阶导数连续，从而其左右导数相等，这样每一个内部节点可产生2个方程，共计 $2(n-1)$ 个。这样就得到 $4n$ 个待定参数所满足的 $2n+2(n-1)=4n-2$ 个方程，为保证待定参数的唯一性，还差两个方程。为此，常用的方法是对边界节点除函数值外附加要求，这就是所谓的边界条件。根据实际问题的不同，三次样条插值常用到下列三类边界条件。

（i）m边界条件：$S'(a)=y'_0, S'(b)=y'_n$，即给定端点处的一阶导数值。由这种边界条件建立的样条插值函数称为 $g(x)$ 的完备三次样条插值函数。特别地，$y'_0=y'_n=0$ 时，样条曲线在端点处呈水平状态。

如果y'_0、y'_n不知道，我们可以要求$S'(x)$在端点处与y'_0、y'_n近似相等，即以x_0,x_1,x_2,x_3为节点作一个三次拉格朗日插值多项式$L_a(x)$，以$x_n,x_{n-1},x_{n-2},x_{n-3}$作一个三次拉格朗日插值多项式$L_b(x)$，使得

$$S'(a)=N'_a(a), S'(b)=N'_b(b).$$

由这种边界条件建立的三次样条插值函数称为$g(x)$的拉格朗日三次样条插值函数。

（ii） M边界条件：$S''(a)=y''_0, S''(b)=y''_n$，即给定端点处的二阶导数值。特别地$y''_0=y''_n=0$时，称为自然边界条件。

（iii） 周期边界条件：当$y=g(x)$是以$b-a=x_n-x_0$为周期的周期函数时，要求$S(x)$也是周期函数，故端点处要满足$S'(a+0)=S'(b-0), S''(a+0)=S''(b-0)$，此条件称为周期条件。

给出任一种边界条件，都可以得到两个独立的方程。这样$4n$个独立方程可唯一确定三次样条插值函数的$4n$个参数，从而三次样条插值函数在理论上是唯一确定的。

三次样条插值函数的表达式通常有两种方式，其具体的表现形式与计算方式可查阅相关资料。

当$g(x)$在$[a,b]$上连续时，满足上述三类边界条件之一的三次样条插值函数$S(x)$有良好的收敛性，即

$$\lim_{h\to 0} S(x)=g(x),\quad x\in[a,b],$$

其中$h=\max\limits_{1\le j\le n}\{x_j-x_{j-1}\}$。而且对于满足m边界条件或M边界条件的三次样条插值函数$S(x)$，当$g(x)$在$[a,b]$上四阶导数连续时有误差估计

$$|g^{(l)}(x)-S^{(l)}(x)|\le c_l M_4 h^{4-l},\quad l=0，1，2，\quad x\in[a,b],$$

其中$c_0=\dfrac{5}{384}$，$c_1=\dfrac{1}{24}$，$c_2=\dfrac{3}{8}$，$M_4=\max\limits_{a\le x\le b}\{|g^{(4)}(x)|\}$。

2 三次样条插值的Matlab实现

在Matlab中数据点称之为断点。如果三次样条插值没有边界条件，最常用的方法，就是采用非扭结（not-a-knot）条件。这个条件强迫第1个和第2个三次多项式的三阶导数相等。对最后一个和倒数第2个三次多项式也做同样地处理。

Matlab中三次样条插值也有现成的函数：

y=interp1(x0,y0,x,'spline');

y=spline(x0,y0,x);

pp=csape(x0,y0,conds),

pp=csape(x0,y0,conds,valconds)，y=ppval(pp,x)。

其中x0,y0是已知数据点，x是插值点，y是插值点的函数值。

对于三次样条插值，我们提倡使用函数csape，csape的返回值是pp形式，要求插值点的函数值，必须调用函数ppval。

pp=csape(x0,y0)：使用默认的边界条件，即Lagrange边界条件。

pp=csape(x0,y0,conds,valconds)中的conds指定插值的边界条件，其值可为：

'complete'　边界为一阶导数，一阶导数的值在valconds参数中给出，若忽略valconds参数，则按缺省情况处理。

'not-a-knot'　非扭结条件

'periodic'　周期条件

'second'　边界为二阶导数，二阶导数的值在valconds参数中给出，若忽略valconds参数，二阶导数的缺省值为[0, 0]。

'variational'　设置边界的二阶导数值为[0,0]。

对于一些特殊的边界条件，可以通过conds的一个1×2矩阵来表示，conds元素的取值为0，1，2。

conds(i)=j的含义是给定端点i的j阶导数，即conds的第一个元素表示左边界的条件，第二个元素表示右边界的条件，conds=[2,1]表示左边界是二阶导数，右边界是一阶导数，对应的值由valconds给出。

详细情况请使用帮助help csape。

例1　机床加工

待加工零件的外形根据工艺要求由一组数据(x,y)给出（在平面情况下），用程控铣床加工时每一刀只能沿x方向和y方向走非常小的一步，这就需要从已知数据得到加工所要求的步长很小的(x,y)坐标。

表中给出的x,y数据位于机翼断面的下轮廓线上，假设需要得到x坐标每改变0.1时的y坐标。试完成加工所需数据，画出曲线，并求出$x=0$处的曲线斜率和$13\le x\le 15$范围内y的最小值。

x	0	3	5	7	9	11	12	13	14	15
y	0	1.2	1.7	2.0	2.1	2.0	1.8	1.2	1.0	1.6

要求用Lagrange、分段线性和三次样条三种插值方法计算。

解　编写以下程序：

```
x0=[0   3   5   7   9   11   12   13   14   15];
y0=[0  1.2  1.7  2.0  2.1  2.0  1.8  1.2   1.0  1.6];
x=0:0.1:15;
y1=lagrange(x0,y0,x);   %前面编写的拉格朗日插值函数
y2=interp1(x0,y0,x);
y3=interp1(x0,y0,x,'spline');
pp1=csape(x0,y0);
```

```
y4=ppval(pp1,x);
pp2=csape(x0,y0,'second');
y5=ppval(pp2,x);
[x',y1',y2',y3',y4',y5']
subplot(2,2,1)
plot(x0,y0,'+',x,y1)
title('Lagrange')
subplot(2,2,2)
plot(x0,y0,'+',x,y2)
title('Piecewise linear')
subplot(2,2,3)
plot(x0,y0,'+',x,y3)
title('Spline1')
subplot(2,2,4)
plot(x0,y0,'+',x,y4)
title('Spline2')
dx=diff(x);
dy=diff(y3);
dy_dx=dy./dx;
dy_dx0=dy_dx(1)
ytemp=y3(131:151);
ymin=min(ytemp);
index=find(y3==ymin);
xmin=x(index);
[xmin,ymin]
```

计算结果略。

可以看出，拉格朗日插值的结果根本不能应用，分段线性插值的光滑性较差（特别是在 $x=14$ 附近弯曲处），建议选用三次样条插值的结果。

五　一维插值总结

插值函数一般是已知函数的线性组合或者称为加权平均。在已知数据点较少时，插值技术在工程实践和科学实验中有着广泛而又十分重要的应用。例如在信息技术中的图像重建、图像放大过程中为避免图像失真、扭曲而增加的插值补点，建筑工程的外观设计，化学工程试验数据与模型分析，天文观测数据、地理信息数据的处理，社会经济现象的统计分析等方面，插值技术的应用是不可或缺的。

插值技术（或方法）远不止这里所介绍的这些，但在解决实际问题时，对于一位插值问题而言，前面介绍的插值方法已经足够了。剩下的问题关键在于什么情况下使用、怎样使用和使用何种插值方法的选择上。

拉格朗日插值函数在整个插值区间上有统一的解析表达式，其形式关于节点对称，光滑性好。但缺点同样明显，这主要体现在高次插值收敛性差（龙格现象）；增加节点时前期计算作废，导致计算量大；一个节点函数值的微小变化（观测误差存在）将导致整个区间上插值函数都发生改变，因而稳定性差等几个方面。因此拉格朗日插值法多用于理论分析，在采用拉格朗日插值方法进行插值计算时通常选取 $n<7$ 。

分段线性插值函数（仅连续）与三次样条插值函数（二阶导数连续）虽然光滑性差，但他们都克服了拉格朗日插值函数的缺点，不仅收敛性、稳定性强，而且方法简单实用，计算量小。因而应用十分广泛。

六　二维插值简介

前面讲述的都是一维插值，其特点是节点为一维实数组，插值函数是一元函数（曲线）。若节点是二维数组，则插值函数就是二元函数，即曲面。例如在某区域测量了若干点（节点）的高程（节点值），为了画出该区域较精确的等高线图，就要先插入更多的点（插值点），计算这些点的高程（插值），这就是一个二维插值问题。

二维插值问题的数学描述为：已知二元函数 $g(x,y)$ 在某矩形区域 $R=[a,b]\times[c,d]$ 上互异节点 (x_i,y_j) 的函数值 z_{ij} ，如何求出在 R 上任一点 (x,y) 处的函数值 $g(x,y)$ 的近似值。

二维插值方法的基本思想是：根据已知数据点 (x_i,y_j,z_{ij}) ，构造一个具有一定光滑性、简单易于计算的函数 $z=f(x,y)$ （已知曲面）作为 $z=g(x,y)$ （未知曲面）的近似，使曲面 $z=f(x,y)$ 通过已知数据点，即 $f(x_i,y_j)=z_{ij}$ 。然后计算点 (x,y) 的函数值 $f(x,y)$ 作为 $g(x,y)$ 的近似值。

常见的二维插值可分为网格节点插值和散乱节点插值。其中网格节点插值适用于节点比较规范的情况，即在包含所给节点的矩形区域内，节点由两族平行于坐标轴的直线的交点所组成。散乱节点插值适用于一般的节点，多用于节点不太规范（即节点为两族平行于坐标轴的直线的部分交点）的情况。

1 网格节点插值法

网格节点插值问题可简述为：已知数据点 (x_i,y_j,z_{ij}) ，其中 i =1，2，…， m ， j =1，2，…， n ， $a=x_1<x_2<\cdots<x_m=b$ ， $c=y_1<y_2<\cdots<y_n=d$ ，求 R 上任一插值点 (x,y) $(\neq(x_i,y_j))$ 处的插值 z 。

网格节点插值法主要有以下几种形式：

（1）　分片线性插值

分片线性插值即是在以 (x_i,y_j) ， (x_{i+1},y_j) ， (x_{i+1},y_{j+1}) ， (x_i,y_{j+1}) 为顶点的小矩形 R_{ij} 上作线性插值。其分片插值函数为：

当 $x_i \le x \le x_{i+1}$，$y_j \le y \le \dfrac{y_{j+1}-y_j}{x_{i+1}-x_i}(x-x_i)+y_j$ 时，

$$f(x,y)=z_{ij}+(z_{i+1j}-z_{ij})(x-x_i)+(z_{i+1j+1}-z_{i+1j})(y-y_j),$$

当 $x_i \le x \le x_{i+1}$，$\dfrac{y_{j+1}-y_j}{x_{i+1}-x_i}(x-x_i)+y_j \le y \le y_{j+1}$ 时，

$$f(x,y)=z_{ij}+(z_{i+1j+1}-z_{ij+1})(x-x_i)+(z_{ij+1}-z_{ij})(y-y_j)。$$

显然，分片线性插值函数在 R 上连续。

(2) 分片双线性插值

分片双线性插值曲面由一片一片空间二次曲面构成，其分片函数表达式为：

$$f(x,y)=(Ax+B)(Cx+D),\quad (x,y)\in R_{ij},$$

其中四个待定系数可由 R_{ij} 四个顶点值唯一确定。但分片双线性插值曲面在 R 上不连续。

(3) 分片双三次样条插值

分片双三次样条插值函数在每一片（即小矩形 R_{ij} ）上的表达式为

$$f(x,y)=(A_1+A_2x+A_3x^2+A_4x^3)(B_1+B_2x+B_3x^2+B_4x^3),$$

其中待定系数 A_i，$B_i(i=1,2,3,4)$ 可由 R_{ij} 四个顶点的值及插值函数 $f(x,y)$ 在 x 和 y 方向的光滑性（即 $f_x^{'}$、$f_{xx}^{''}$、$f_y^{'}$、$f_{yy}^{''}$ 连续）和相应的边界条件所唯一确定。

2 散乱节点插值

散乱节点插值问题可简述为：已知函数 $z=f(x,y)$ 在矩形域 R 上散乱分布的 n 个互异节点 (x_i,y_j) 处的函数值 z_{ij}，求 R 上任一插值点 $(x,y)\ (\ne(x_i,y_j))$ 的插值 z 。

常用的方法是反距离加权平均法，或称Shepard方法。其基本思想是，在点 (x,y) $(\ne(x_i,y_j))$，定义其插值函数的函数值为节点处函数值按 (x,y) 与节点距离的某种形式反比作为权重的加权平均。例如若记 $r_{ij}=\sqrt{(x-x_i)^2+(y-y_j)^2}$，则插值函数（曲面）可定义为

$$f(x,y)=\begin{cases} z_{ij}, & r_{ij}=0, \\ \sum_{i,j} W_{ij}(x,y)z_{ij}, & r_{ij}\ne 0, \end{cases}$$

其中 $W_{ij}(x,y)=\dfrac{1}{r_{ij}^2}\Big/\sum_{i,j}\dfrac{1}{r_{ij}^2}$。

这样定义的插值曲面是全局相关的，对曲面上任一点作数值计算都要涉及到全体已知数据，因而在已知数据量大的情况下计算量相当大，而且插值曲面在节点 (x_i,y_j) 处不光滑。虽然如此，由于该方法思想简单，故有种种改进方法。

3 二维插值的Matlab实现

Matlab中有一些计算二维插值的程序。如

```
z=interp2(x0,y0,z0,x,y,'method')
```

其中x0,y0分别为m维和n维向量，表示节点，z0为$n \times m$维矩阵，表示节点值，x,y为一维数组，表示插值点，x与y应是方向不同的向量，即一个是行向量，另一个是列向量，z为矩阵，表示得到的插值，'method'的用法同上面的一维插值。

如果是三次样条插值，可以使用命令

pp=csape({x0,y0},z0,conds,valconds)，y=ppval(pp,{x,y})

其中x0,y0分别为m维和n维向量，z0为$m \times n$维矩阵，具体使用方法同一维插值。

例2 在一丘陵地带测量高程，x和y方向每隔100米测一个点，得高程如下表，试拟合一曲面，确定合适的模型，并由此找出最高点和该点的高程。

x y	100	200	300	400	500
100	636	697	624	478	450
200	698	712	630	478	420
300	680	674	598	412	400
400	662	626	552	334	310

解 编写程序如下：

```
clear,clc
x=100:100:500;
y=100:100:400;
z=[636     697     624     478    450
   698     712     630     478    420
   680     674     598     412    400
   662     626     552     334    310];
pp=csape({x,y},z')
cz1=fnval(pp,{100:10:500,100:10:400})
cz2=interp2(x,y,z,100:10:500,[100:10:400]','spline')
```

§3 数据拟合

在科学计算中经常要建立实验数据的数学模型。给定函数的实验数据，需要用比较简单和合适的函数来逼近（或拟合）实验数据。这种逼近的特点是：

(a) 适度的精度是需要的；

(b) 实验数据有小的误差；

(c) 对于某些问题，可能有某些特殊的信息能够用来选择实验数据的数学模型。

逼近离散数据的基本方法就是曲线拟合，常采用最小二乘拟合。

曲线拟合问题的数学描述是，已知一组（二维）数据(x_i, y_i)，$i = 1,2,\cdots,n$（即平面上的n个点(x_i, y_i)，$i = 1,2,\cdots,n$），x_i互不相同。寻求一个函数（曲线）$y = f(x)$，使$f(x)$在某种准则下与所有数据点最为接近，即曲线拟合得最好。

最小二乘拟合分为线性最小二乘拟合和非线性最小二乘拟合。

一　**线性**最小二乘拟合

线性最小二乘法是解决曲线拟合问题最常用的方法，基本思路是，令

$$f(x)=a_1r_1(x)+a_2r_2(x)+\cdots+a_mr_m(x)\text{，}\qquad (10.4)$$

其中$\{r_k(x)\}_{k=1}^m\ (m<n)$是一组事先选定的线性无关的函数，$\{a_k\}_{k=1}^m$是一组待定系数。寻求系数$\{a_k\}_{k=1}^m$使得$y_i$与$f(x_i)$的距离$d_i$（$i=1,2,\cdots,n$）的平方和最小。这种准则称为最小二乘准则，其求系数$\{a_k\}_{k=1}^m$的方法称为线性最小二乘拟合方法。

1 系数$\{a_k\}_{k=1}^m$的求法

若记

$$J(a_1,\cdots,a_m)=\sum_{i=1}^n d_i^2=\sum_{i=1}^n[f(x_i)-y_i]^2\text{，}\qquad (10.5)$$

则J为$a_1,\cdots,a_m$的二次函数。由数学分析（或高等数学）的极值理论，J达到最小的充分必要条件是$a_1,\cdots,a_m$满足$\dfrac{\partial J}{\partial a_k}=0\quad(k=1,\cdots,m)$。于是得到求使$J$达到最小的$a_1$，…，$a_m$的方法是求解线性方程组（称为法方程组）

$$\sum_{i=1}^n r_j(x_i)[\sum_{k=1}^m a_kr_k(x_i)-y_i]=0,\quad (j=1,\cdots,m)\text{，}$$

即求解线性方程组

$$\sum_{k=1}^m a_k[\sum_{k=1}^n r_j(x_i)r_k(x_i)]=\sum_{i=1}^n r_j(x_i)y_i,\quad (j=1,\cdots,m)\text{。}\qquad (10.6)$$

若记

$$R=\begin{bmatrix} r_1(x_1) & \cdots & r_m(x_1)\\ \vdots & \vdots & \vdots\\ r_1(x_n) & \cdots & r_m(x_n)\end{bmatrix}_{n\times m}\text{，}\quad A=[a_1,\cdots,a_m]^T\text{，}\quad Y=(y_1,\cdots,y_n)^T\text{，}$$

则方程组（10.6）可表示为

$$R^TRA=R^TY\text{。}\qquad (10.7)$$

由于当$\{r_1(x),\cdots,r_m(x)\}$线性无关时，R列满秩，R^TR可逆，所以方程组（10.7）有唯一解

$$A=(R^TR)^{-1}R^TY\text{。}$$

用（10.4）作线性最小二乘拟合的误差通常考虑以下两种形式：

最小平方误差：$\delta_1=(\sum_{i=1}^n(y_i-f(x_i)))^{1/2}$；

最大偏差：$\delta_2=\max\limits_{1\le i\le n}|y_i-f(x_i)|$。

2 函数组$\{r_k(x)\}_{k=1}^m$的选取

面对一组数据$(x_i,y_i),i=1,2,\cdots,n$，用线性最小二乘法作曲线拟合时，首要的、也是关键

的一步是恰当地选取$r_1(x),\cdots,r_m(x)$。如果通过机理分析，能够知道y与x之间应该有什么样的函数关系，则$r_1(x),\cdots,r_m(x)$容易确定。若无法知道y与x之间的关系，通常可以将数据$(x_i,y_i),i=1,2,\cdots,n$作图，直观地判断应该用什么样的曲线去作拟合。人们常用的曲线有

（i）直线　$y=a_1x+a_2$；

（ii）多项式　$y=a_1x^m+\cdots+a_mx+a_{m+1}$（一般$m=2,3$，不宜太高）；

（iii）双曲线（一支）　$y=\frac{a_1}{x}+a_2$，拟合前作变量替换$t=1/x$求解a_1,a_2较简单；

（iv）指数曲线　$y=a_1e^{a_2x}$，拟合前作变量代换$z=\ln y$，$t=1/x$，则指数曲线$y=a_1e^{a_2x}$转化为关于$\ln a_1,a_2$的线性函数$z=\ln a_1+a_2t$，这样做求解a_1,a_2较简单。

在实际计算过程中，面对一组已知数据，到底用什么样的曲线拟合最好，可以在直观判断的基础上，选几种曲线分别拟合，然后比较，看哪条曲线的最小二乘指标J最小。

二　非线性最小二乘拟合

非线性最小二乘法是假设$f(x)$是待定系数$\{a_k\}_{k=1}^m$的任意非线性函数，在最小二乘准则下求其系数$\{a_k\}_{k=1}^m$。

例如上述人们常用的双曲线和指数曲线就是非线性最小二乘拟合中最常用的非线性函数，只不过在上面使用中将它们转变成线性最小二乘拟合方法。

对于给定的实验数据，通常应根据实验数据的走向、趋势选择合适的数学模型，即拟合函数。例如当实验数据具有单调性和凸性时，可选择下述适当的数学模型$y=f(x)$来拟合实验数据。

$f(x)=ae^{bx}$，$f(x)=ae^{b/x}$，$f(x)=ax^b$，$f(x)=a+b/x$

等，其中a、b为参数。

在有可能的情况下，一般将非线性拟合函数转化为线性拟合函数求解，这一方面是如此求解简单，另一方面也是因为一般情况下求解法方程组$\frac{\partial J}{\partial a_k}=0\quad(k=1,\cdots,m)$得到的$(a_1,a_2,\cdots,a_k)$通常仅是$J$的驻点，不一定是极值点。也可以直接解$J$极小化问题。

三　最小二乘拟合法的Matlab实现

1 解方程组方法

在上面的记号下，

$J(a_1,\cdots,a_m)=\|RA-Y\|^2$。

Matlab中的线性最小二乘的标准型为

$$\operatorname*{Min}_A \|RA-Y\|_2^{\,2},$$

命令为 $A=R\backslash Y$。

例3 用最小二乘法求一个形如$y=a+bx^2$的经验公式，使它与下表所示的数据拟合。

x	19	25	31	38	44

y	19.0	32.3	49.0	73.3	97.8

解　编写程序如下

```
x=[19     25    31     38    44]';
y=[19.0   32.3   49.0   73.3   97.8]';
r=[ones(5,1),x.^2];
ab=r\y
x0=19:0.1:44;
y0=ab(1)+ab(2)*x0.^2;
plot(x,y,'o',x0,y0,'r')
```

2 线性最小二乘拟合（多项式拟合）方法

在线性最小二乘拟合中，用的较多的是多项式拟合。如果取 $\{r_1(x),\cdots,r_{m+1}(x)\}=\{1,x,\cdots,x^m\}$，即用 m 次多项式拟合给定数据，则Matlab中有现成的函数

a=polyfit(x0,y0,m)，

其中输入参数x0,y0为要拟合的数据，m为拟合多项式的次数，输出参数a为拟合多项式 $y=a_m x^m+\cdots+a_1 x+a_0$ 系数a=[a_m, …, a_1, a_0]。

多项式在x处的值y可用下面的函数计算

y=polyval(a,x)。

例4 某乡镇企业1990-1996年的生产利润如下表：

年份	1990	1991	1992	1993	1994	1995	1996
利润（万元）	70	122	144	152	174	196	202

试预测1997年和1998年的利润。

解　作已知数据的的散点图，

```
x0=[1990  1991  1992  1993  1994  1995  1996];
y0=[70   122   144   152   174   196   202];
plot(x0,y0,'*')
```

发现该乡镇企业的年生产利润几乎直线上升。因此，我们可以用 $y=a_1x+a_0$ 作为拟合函数来预测该乡镇企业未来的年利润。编写程序如下：

```
x0=[1990  1991  1992  1993  1994  1995  1996];
y0=[70   122   144   152   174   196   202];
a=polyfit(x0,y0,1)
y97=polyval(a,1997)
y98=polyval(a,1998)
```

求得 $a_1=20$，$a_0=-4.0705\times10^4$，1997年的生产利润y97=233.4286，1998年的生产利

润y98=253.9286。

3 非线性最小二乘拟合

Matlab的优化工具箱中提供了两个求非线性最小二乘拟合的函数：curvefit和leastsq。使用这两个命令时，都要先建立M文件fun.m，但它们定义$f(x)$的方式是不同的。

1　curvefit

设已知 $xdata=(xdata_1, xdata_2, \cdots, xdata_n)$, $ydata=(ydata_1, ydata_2, \cdots, ydata_n)$, curvefit用以求含参量x（向量）的向量值函数

$$F(x, xdata)=(f(x, data_1), \cdots, f(x, xdata_n))^T$$

中的参变量x（向量），使得

$$\sum_{i=1}^{n} (F(x, xdata_i)-ydata_i)^2$$

最小。

输入格式为：

（1）x=curvefit('fun',x0,xdata,ydata)；

（2）x=curvefit('fun',x0,xdata,ydata,options)；

（3）x=curvefit('fun',x0,xdata,ydata,options, 'grad')；

（4）[x，options]=curvefit('fun',x0,xdata,ydata,…)；

（5）[x，options,funval]=curvefit('fun',x0,xdata,ydata,…)；

（6）[x，options,funval,Jacob]=curvefit('fun',x0,xdata,ydata,…).

输出目标函数值格式：f=fun(x,xdata).

其中x0为迭代初值，options为控制参数。

2　leastsq

设已知 $xdata=(xdata_1, xdata_2, \cdots, xdata_n)$, $ydata=(ydata_1, ydata_2, \cdots, ydata_n)$, leastsq 用以求含参量x（向量）的向量值函数

$$f(x)=(f_1(x), f_2(x) \cdots, f_n(x))^T$$

中的参变量x（向量），使得

$$f(x)^2 f(x)= f_1(x)^2+ f_2(x)^2 +\cdots+f_n(x)^2$$

最小。其中

$$f_i(x)=F(x, xdata_i, ydata_i)$$

输入格式为：

（1）x= leastsq ('fun',x0)；

（2）x= leastsq ('fun',x0,options)；

（3）x= leastsq ('fun',x0,options, 'grad')；

（4）[x，options]= leastsq ('fun',x0,…)；

（5）[x，options,funval]= leastsq（'fun',x0,…）;

例5 用下面一组数据拟合函数$c(t)=a+be^{-0.02kt}$中的参数a,b,k。

t_j	100	200	300	400	500	600	700	800	900	1000
$c_j \times 10^3$	4.54	4.99	5.35	5.65	5.90	6.10	6.26	6.39	6.50	6.59

解　该问题即阶最优化问题

$$\min F(a,b,k)=\sum_{j=1}^{10}[a+be^{-0.02kt_j}-c_j]^2$$

。

1 用命令curvefit。此时

F(x，tdata)=(a+b$e^{-0.02kt_1}$，…，a+b$e^{-0.02kt_{10}}$)^T，x=(a，b，k)

（1） 编写M文件curvefun1.m

```
function f=curvefun1(x,tdata)
f=x(1)+x(2)*exp(-0.02*x(3)*tdata)   %其中x(1)=a；x(2)=b；x(3)=k；
```

（2） 输入命令

```
tdata=100:100:1000
cdata=1e-03*[4.54,4.99,5.35,5.65,5.90,6.10,6.26,6.39,6.50,6.59];
x0=[0.2,0.05,0.005];
x=curvefit('curvefun1',x0,tdata,cdata)
f=curvefun1(x,tdata)
```

（3）运算结果为：

```
x=0.0070   -0.0030   0.1012
f=
Columns 1 through 7
0.0045   0.0050   0.0054   0.0057   0.0059   0.0061   0.0063
Columns 8 through 10
0.0064   0.0065   0.0066
```

即拟合得a=0.0070，b=-0.0030，k=0.0066。

2 用命令leastsq。此时

f(x)=F(x,tdata,cdata)=(a+b$e^{-0.02kt_1}$-c_1，…，a+b$e^{-0.02kt_{10}}$-c_{10})^T

x=(a，b，k)

（1） 编写M文件curvefun2.m

```
function f=curvefun2(x)
tdata=100:100:1000;
```

```
cdata=1e-03*[4.54,4.99,5.35,5.65,5.90,6.10,6.26,6.39,6.50,6.59];
  f=cdata-x(1)-x(2)*exp(-0.02*x(3)*tdata)  %其中x(1)=a；x(2)=b；x(3)=k；
```

（2） 输入命令

```
x0=[0.2,0.05,0.005];
x=leastsq('curvefun2',x0)
f=curvefun2(x)
```

（3）运算结果为：

```
x=0.0070   -0.0030   0.1012
f=1.0e-005*
Columns 1 through 7
0.0221   0.2081   -0.3933   -0.2872   0.2973   0.3561   0.0693
Columns 8 through 10
-0.2327   -0.0970   0.0296
```

可以看出，两个命令的计算结果是相同的。

§4 函数的最小二乘逼近

在科学计算中我们还常遇到已知函数的逼近问题。由于电子计算机只能做算术运算，因而在计算机上计算数学函数（如　等在有限区间上的计算）必须用其它简单的函数（如多项式或有理分式）来逼近，且用该函数代替原来精确的数学函数进行计算。这种函数逼近的特点是：

（a） 要求是高精度逼近；

（b） 要快速计算（计算量越小越好）。

前面讲的曲线拟合是已知一组离散数据 $\{(x_i, y_i), i=1,\cdots,n\}$，选择一个较简单的函数 $f(x)$，如多项式，在一定准则如最小二乘准则下，最接近这些数据。

如果已知一个较为复杂的连续函数 $y(x), x\in[a,b]$，要求选择一个较简单的函数 $f(x)$，在一定准则下最接近 $f(x)$，就是所谓函数逼近。

与曲线拟合的最小二乘准则相对应，函数逼近常用的一种准则是最小平方逼近，即

$$J=\int_a^b [f(x)-y(x)]^2\,dx \tag{10.8}$$

达到最小。与曲线拟合一样，选一组函数 $\{r_k(x), k=1,\cdots,m\}$ 构造 $f(x)$，即令

$$f(x)=a_1 r_1(x)+\cdots+a_m r_m(x)$$

代入（11）式，求 $a_1,\cdots,a_m$ 使 J 达到极小。利用极值必要条件可得

$$\begin{bmatrix} (r_1,r_1) & \cdots & (r_1,r_m) \\ \vdots & \vdots & \vdots \\ (r_m,r_1) & \cdots & (r_m,r_m) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} (y,r_1) \\ \vdots \\ (y,r_m) \end{bmatrix}, \tag{10.9}$$

这里 $(g,h)=\int_a^b g(x)h(x)dx$ 。当方程组（10.9）的系数矩阵非奇异时，有唯一解。

最简单的当然是用多项式逼近函数，即选 $r_1(x)=1$ ，$r_2(x)=x$ ，$r_3(x)=x^2$ ，…。并且如果能使 $\int_a^b r_i(x)r_j(x)dx=0$ ，$(i\neq j)$，方程组（10.9）的系数矩阵将是对角阵，计算大大简化。满足这种性质的多项式称正交多项式。

勒让得（Legendre）多项式是在 $[-1,1]$ 区间上的正交多项式，它的表达式为

$$P_0(x)=1,\quad P_k(x)=\frac{1}{2^k k!}\frac{d^k}{dx^k}(x^2-1)^k, k=1,2,\cdots$$

可以证明

$$\int_{-1}^1 P_i(x)P_j(x)dx=\begin{cases}0, & i\neq j\\ \dfrac{2}{2i+1}, & i=j\end{cases}$$

$$P_{k+1}(x)=\frac{2k+1}{k+1}xP_k(x)-\frac{k}{k+1}P_{k-1}(x),\quad k=1,2,\cdots$$

常用的正交多项式还有第一类切比雪夫（Chebyshev）多项式

$$T_n(x)=\cos(n\arccos x),\quad (x\in[-1,1], n=0,1,2,\cdots)$$

和拉盖尔（Laguerre）多项式

$$L_n(x)=e^x\frac{d^n}{dx^n}(x^n e^{-x}),\quad (x\in[0,+\infty), n=0,1,2,\cdots)$$

。

例6 求 $f(x)=\cos x, x\in\left[-\frac{\pi}{2},\frac{\pi}{2}\right]$ 在 $H=\mathrm{Span}\{1,x^2,x^4\}$ 中的最佳平方逼近多项式。

解　编写程序如下：

```
syms x
base=[1,x^2,x^4];
y1=base.'*base
y2=cos(x)*base.'
r1=int(y1,-pi/2,pi/2)
r2=int(y2,-pi/2,pi/2)
a=r1\r2
xishu1=double(a)
digits(8),xishu2=vpa(a)
```

求得xishu1=0.9996　　-0.4964　　0.0372，即所求的最佳平方逼近多项式为

$$y=0.9996-0.4964x^2+0.0372x^4.$$

§5 回归分析

回归分析是处理很难用一种精确方法表示出来的变量之间关系的一种数学方法，它是最常用的数理统计方法，能解决预测、控制、生产工艺优化等问题。它在工农业生产和科学研究各个领域中均有广泛的应用。

回归分析一般分为线性回归分析和非线性回归分析。本节着重介绍线性回归分析的基本结论及其在Matlab中的相应命令。线性回归分析是两类回归分析中较简单的一类，也是应用较多的一类。

一 一元线性回归分析

针对一组（二维）数据(x_i, y_i)，$i = 1,2,\cdots,n$（其中x_i互不相同），其最简单的数据拟合形式为寻求直线$y = \beta_0 + \beta_1 x$，使$\beta_0 + \beta_1 x$在最小二乘准则下与所有数据点最为接近。但由于随机观测误差的存在，满足上述数据点的直线应该是

$$y = \beta_0 + \beta_1 x + \varepsilon, \tag{10.10}$$

其中x, y是准确的，β_0, β_1是两个未知参数，ε是均值为零的随机观测误差，具有不可观测性，可以合理地假设这种观测误差服从正态分布。于是我们得到一元线性回归模型为

$$\begin{cases} y = \beta_0 + \beta_1 x + \varepsilon, \\ E\varepsilon = 0, D\varepsilon = \sigma^2, \end{cases} \tag{10.11}$$

其中σ未知，固定的未知参数β_0、β_1称为回归系数，自变量x称为回归变量。

(10.10)式两边同时取期望得：$Y = \beta_0 + \beta_1 x$，称为y对x的回归直线方程。

在该模型下，第i个观测值可以看作样本$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$（这些样本相互独立但不同分布，$i = 1,2,\cdots,n$）的实际抽样值，即样本值。

一元线性回归分析的主要任务是：用实验值（样本值）对β_0、β_1和σ作点估计；对回归系数β_0、β_1作假设检验；在$x = x_0$处对y作预测，并对y作区间估计。

1回归参数β_0、β_1和σ^2估计

假设有n组独立观测值：(x_i, y_i)，$i = 1,2,\cdots,n$，则由(10.11)有

$$\begin{cases} y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \\ E\varepsilon_i = 0, D\varepsilon_i = \sigma^2, \end{cases}$$

其中ε_i，$i = 1,2,\cdots,n$相互独立。记

$$Q = Q(\beta_0, \beta_1) = \sum_{i=1}^n \varepsilon_i^{\ 2} = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 。$$

称$Q(\beta_0, \beta_1)$为偏离真实直线的偏差平方和。由最小二乘法得到的估计$\beta_i (i = 0,1)$称为β_i的最小二乘估计，其中

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad \hat{\beta}_1 = \frac{\hat{y} - \bar{x}\bar{y}}{\hat{x} - \bar{x}^2}, \quad \bar{x} = \frac{1}{n}\sum_{i=1}^n x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^n y_i, \quad \hat{x} = \frac{1}{n}\sum_{i=1}^n x_i^2, \quad \hat{y} = \frac{1}{n}\sum_{i=1}^n x_i y_i 。$$

（经验）回归方程为

$$Y=\beta_0+\beta_1 x=\bar{y}+\beta_1(x-\bar{x})。 \quad (10.12)$$

这样我们得到 $\beta_i(i=0,1)$ 的无偏估计 $\beta_i(i=0,1)$，其中 $\beta_i(i=0,1)$ 服从正态分布

$$\hat{\beta}_0 \sim N(\beta_0,\sigma^2(1/n+\bar{x}^2\Big/\sum_{i=1}^{n}(x_i-\bar{x})^2),\quad \hat{\beta}_1 \sim N(\beta_1,\sigma^2\Big/\sum_{i=1}^{n}(x_i-\bar{x})^2),$$

且对每组 (x_i,y_i)，可以求出拟合值 Y_i 以及残差 y_i-Y_i，易知

$$\sum_{i=1}^{n}(y_i-Y_i)=0。$$

这说明残差之和为零。

若记

$$Q_e=Q(\hat{\beta}_0,\hat{\beta}_1)=\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)^2=\sum_{i=1}^{n}(y_i-Y_i)^2,$$

则称 Q_e 为残差平方和或剩余平方和。在模型(10.11)下，易知，$EQ_e=(n-2)\sigma^2$，常记

$$\hat{\sigma}_e^2=Q_e/(n-2)。$$

称 $\hat{\sigma}_e^2$ 为剩余方差（残差的方差），它是 σ^2 的无偏估计，且有如下分布

$$\frac{1}{\sigma^2}Q_e \sim \chi^2(n-2),\quad \frac{n-2}{\sigma^2}\hat{\sigma}_e^2 \sim \chi^2(n-2)。$$

$\hat{\sigma}_e^2$ 分别与 β_0、β_1 独立，$\hat{\sigma}_e$ 称为剩余标准差，显然，$\hat{\sigma}_e$ 越接近于0，说明线性回归方程(10.12)越显著。

2 模型的假设、预测、控制

1回归方程的显著性检验

在实际问题中，因变量 y 与自变量 x 之间是否有线性关系(10.10)只是一种假设，在求出回归方程之后，还必须对这种回归方程同实际观测数据拟合的效果进行检验。

由(10.10)可知，$|\beta_1|$越大，y 随 x 变化的趋势就越明显；反之，$|\beta_1|$越小，y 随 x 变化的趋势就越不明显。特别当 $\beta_1=0$时，则认为 y 与 x 之间不存在线性关系，当 $\beta_1\neq 0$时，则认为 y 与 x 之间有线性关系(10.10)。因此，问题归结为对假设

$$H_0:\beta_1=0;\quad H_1:\beta_1\neq 0$$

进行检验。假设 $H_0:\beta_1=0$ 被拒绝，则回归显著，认为 y 与 x 之间存在线性关系，所求的线性回归方程有意义；否则回归不显著，y 与 x 的关系不能用一元线性回归模型来描述，所得的回归方程也无意义。此时，可能有如下几种情况：

（i）x 对 y 没有显著影响，此时应丢掉变量 x；

（ii）x 对 y 有显著影响，但这种影响不能用线性关系来表示，应该用非线性回归；

（iii）除 x 之外，还有其他不可忽略的变量对 y 有显著影响，从而削弱了 x 对 y 的影响。此时应用多元线性回归模型。因此，在接受 H_0 的同时，需要进一步查明原因以便分别处理。

下面介绍两种检验方法，分别是

(a)F检验法

对样本方差 $S=\sum_{i=1}^{n}(y_i-\bar{y})^2$ 进行分解，有

$$S=Q_e+U,\quad U=\sum_{i=1}^{n}(Y_i-\bar{y})^2。$$

上式中的 Q_e 是由实际观测值没有落在回归直线上引起的（否则为零），U 是由回归直线引起的。因此，U 越大，Q_e 就越小，表示 y 与 x 的线性关系就越显著；否则，U 越小，Q_e 就越大，表示 y 与 x 的线性关系就越不显著。这样我们就找到了一种判别回归直线拟合程度好坏的方法：如果 U/S 接近于1，即 U/Q_e 较大时，则对拟合效果感到满意。

易知，当 H_0 成立时，有 $U/\sigma^2 \sim \chi^2(1)$ 且 Q_e 与 U 独立。再由F分布有

$$F=\frac{U}{Q_e/(n-2)}=\frac{(n-2)r^2}{1-r^2}\sim F(1,n-2),\quad r=\sqrt{\frac{U}{U+Q_e}},$$

其中 r 称为相关系数。对给定的显著水平 α，有置信水平为 $1-\alpha$ 的临界值 $F_{1-\alpha}(1,n-2)$，从而F检验法的检验准则为：当 $F>F_{1-\alpha}(1,n-2)$ 时，拒绝 H_0；否则就接受 H_0。

（b）t检验法

当 H_0 成立时，由T分布的定义有

$$T=\frac{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\,\hat{\beta}_1}{\hat{\sigma}_e}\sim t(n-2)。$$

因此，对于给定的显著水平 α，用T统计量检验 H_0，有置信水平为 $1-\alpha$ 的临界值 $t_{1-\alpha/2}(1,n-2)$，从而t检验法的检验准则为：

当 $|T|>t_{1-\alpha/2}(n-2)$ 时，拒绝 H_0；否则就接受 H_0。

2回归系数的区间估计

令 $V=\sum_{i=1}^{n}(x_i-\bar{x})^2$，则由

$$(\hat{\beta}_0-\beta_0)\big/(\hat{\sigma}_e\sqrt{1/n+\bar{x}^2/V})\sim t(n-2),\quad (\hat{\beta}_1-\beta_1)\big/(\hat{\sigma}_e\sqrt{V})\sim t(n-2)$$

得到在置信水平 $1-\alpha$ 下，β_0 和 β_1 的置信区间分别为

$$[\hat{\beta}_0-t_{1-\alpha/2}(n-2)\hat{\sigma}_e\sqrt{1/n+\bar{x}^2/V},\ \hat{\beta}_0+t_{1-\alpha/2}(n-2)\hat{\sigma}_e\sqrt{1/n+\bar{x}^2/V}],$$

$$[\hat{\beta}_1-t_{1-\alpha/2}(n-2)\hat{\sigma}_e/\sqrt{V},\ \hat{\beta}_1+t_{1-\alpha/2}(n-2)\hat{\sigma}_e/\sqrt{V}]。$$

再由 $Q_e/\sigma^2\sim\chi^2(n-2)$，还可得在置信水平 $1-\alpha$ 下，σ^2 的置信区间为

$$\left[\frac{Q_e}{\chi^2_{1-\alpha/2}(n-2)},\ \frac{Q_e}{\chi^2_{\alpha/2}(n-2)}\right]。$$

3） 预测与控制

当检验结果拒绝了$H_0:\beta_1=0$，接下来的问题是如何利用回归方程$Y=\beta_0+\beta_1 x$进行预测和控制。预测就是对固定的x值预测相应的y值，控制就是通过控制x的值，以便把y的值控制在制定的范围内。

(a)预测

设y与x满足模型(10.11)，并且通过了假设检验，即检验结果拒绝了$H_0:\beta_1=0$。令x_0表示x的某个固定值，且$y_0=\beta_0+\beta_1 x_0+\varepsilon_0$，$\varepsilon_0 \sim N(0,\sigma^2)$。假设$y_0,y_1,\cdots,y_n$相互独立，则$y_0$的预测值和预测区间如下。

y_0的预测值为y_0的回归值$Y_0=\beta_0+\beta_1 x_0$。它是Ey_0的无偏估计，即

$$EY_0=E(\hat{\beta}_0+\hat{\beta}_1 x_0)=\beta_0+\beta_1 x_0=Ey_0\text{。}$$

给定显著水平α，y_0的置信水平为$1-\alpha$的预测区间为$[Y_0-\delta(x_0)$，$Y_0+\delta(x_0)]$，其中

$$\delta(x_0)=\hat{\sigma}_e t_{1-\alpha/2}(n-2)\sqrt{1+1/n+(x_0-\bar{x})^2/V}\text{。}$$

由上式可知，剩余标准差$\hat{\sigma}_e$越小，预测区间越小，预测值越精确；对于给定的样本观测值和置信水平而言，x_0越靠近$\bar{x}$时，预测精度就越高。

由于x_0的任意性，夹在两曲线

$$Y_1(x_0)=Y_0-\delta(x_0),\quad Y_2(x_0)=Y_0+\delta(x_0)$$

之间的部分就是$y=\beta_0+\beta_1 x+\varepsilon$的置信水平为$1-\alpha$的预测带。特别，当$n$很大且$x$在$\bar{x}$附近取值时，有

$$t_{1-\alpha/2}(n-2)\approx u_{1-\alpha/2},\quad \sqrt{1+1/n+(x_0-\bar{x})^2/V}\approx 1,$$

其中$u_{1-\alpha/2}$是置信水平为$1-\alpha$的标准正态分布的临界值。此时，$\delta(x)\approx\hat{\sigma}_e u_{1-\alpha/2}$，于是$y$的置信水平为$1-\alpha$的预测区间近似为[$\bar{Y}-\hat{\sigma}_e u_{1-\alpha/2}$，$\bar{Y}+\hat{\sigma}_e u_{1-\alpha/2}$]。在这种情况下，预测带是平行于回归直线的两直线之间的部分，预测工作得到了很大的简化。例如，在应用中可简单地取$\alpha=0.05$，则$u_{1-\alpha/2}=u_{0.975}=1.96$；取$\alpha=0.01$，则$u_{1-\alpha/2}=u_{0.995}=2.58$。从而得到置信水平为95%与99%的预测区间分别近似为[$\bar{Y}-1.96\hat{\sigma}_e$，$\bar{Y}+1.96\hat{\sigma}_e$]和[$\bar{Y}-2.58\hat{\sigma}_e$，$\bar{Y}+2.58\hat{\sigma}_e$]。

(b)控制

若要$y=\beta_0+\beta_1 x+\varepsilon$的值以$1-\alpha$的概率落在指定区间$(c,d)$之内，变量$x$应控制在什么范围内的问题就是所谓的控制问题。它是预测问题的反问题。

由预测的讨论知，夹在两曲线$Y_1(x_0)=Y_0-\delta(x_0)$，$Y_2(x_0)=Y_0+\delta(x_0)$之间的部分就是$y=\beta_0+\beta_1 x+\varepsilon$的置信水平为$1-\alpha$的预测带，故若要$y$的观测值以$1-\alpha$的概率落在区间$(c,d)$之内，只要控制$x$满足以下两不等式

$$\bar{Y}-\delta(x)\ge c,\quad \bar{Y}+\delta(x)\le d\text{。}$$

这要求$d-c\ge 2\delta(x)$。若方程$\bar{Y}-\delta(x)=c$，$\bar{Y}+\delta(x)=d$分别有解a,b，则(a,b)就是所求

的x的控制区间。

一般来说，要解出a和b很复杂。若样本容量很大，即n很大且x在$\bar{x}$附近取值时，一个简化的方法就是利用近似预测区间 [$\bar{Y}-\hat{\sigma}_e u_{1-\alpha/2}$，$\bar{Y}+\hat{\sigma}_e u_{1-\alpha/2}$]来进行控制。

注：一元线性回归是后面介绍的多于线性回归的特例，其Matlab实现见本节第五部分。

二　可线性化的一元非线性回归（曲线回归）

在工程技术中，自变量x与因变量y之间有时呈现出非线性（或曲线）关系，这是通常出现两种情况：一种是呈现多项式的关系，这种情况通过变量替换可化为多元线性回归问题给予解决；另一种是呈现出其它非线性关系，通过变量替换可化为一元线性回归问题给予解决。

已知变量x与y的一组（二维）观测数据(x_i, y_i)，$i=0,1,\cdots,n$（其中x_i互不相同），通常做出二维平面上的散点图，若散点的分布不接近于一条直线，通常根据散点图的类型匹配相应的拟合曲线。若匹配曲线为多项式

$$f(x)=\beta_0+\beta_1 x+\cdots+\beta_m x^m \quad (m<n),$$

这是线性回归问题，即关于回归系数β_i（$i=0,1,\cdots,m$）是线性的。此时只需令

$$x_i = x^i,\quad i=1,2,\cdots,m$$

则回归问题化为多元线性回归问题，其解决方法详见下面介绍的多元线性回归问题（另外，多项式回归问题的实现在Matlab中有相应的命令，具体见多元线性回归部分）。若匹配曲线（经验公式）为含参量a,b的非线性曲线，采用的办法是通过变量替换把非线性回归化为线性回归。通常匹配的含参量a,b的非线性曲线有以下六类，具体的替换方法如下：

1 双曲线$\frac{1}{y}=a+\frac{b}{x}$。

作变量替换$u=\frac{1}{x}$，$v=\frac{1}{y}$，得$v=a+bu$。由观测值(x_i, y_i)，$i=0,1,\cdots,n$按$u_i=\frac{1}{x_i}$，$v_i=\frac{1}{y_i}$算出(u_i, v_i)，$i=0,1,\cdots,n$。对u与v利用前面的经验回归直线公式，计算出参数估计值$\hat{\beta}_0$、$\hat{\beta}_1$，则有回归方程

$$\frac{1}{y}=\hat{\beta}_0+\frac{\hat{\beta}_1}{x}。$$

2 幂函数曲线$y=ax^b$，其中$x>0, a>0$。

两边取常用对数：$\lg y=\lg a+b\lg x$，再作代换$u=\lg x$，$v=\lg y$，$A=\lg a$，则幂函数曲线方程就变成直线方程

$$v=A+bu。$$

由观测值(x_i, y_i)，$i=0,1,\cdots,n$按$u_i=\lg x_i$，$v_i=\lg y_i$算出(u_i, v_i)，$i=0,1,\cdots,n$。对u与v利用本节前面的经验回归直线公式，计算出参数估计值$\hat{A}$、$\hat{b}$，又由$\hat{a}=e^{\hat{A}}$得到$\hat{a}$。因此有回归方程

$$y = \hat{a} x^b$$。

3 指数函数曲线 $y = ae^{bx}$，其中 $a > 0$。

两边取常用对数：$\lg y = \lg a + bx$，再作代换 $u = x$，$v = \lg y$，$A = \lg a$，则指数函数曲线方程就变成直线方程

$$v = A + bu$$。

由观测值 (x_i, y_i)，$i = 0,1,\cdots,n$ 按 $u_i = x_i$，$v_i = \lg y_i$ 算出 (u_i, v_i)，$i = 0,1,\cdots,n$。对 u 与 v 利用本节前面的经验回归直线公式，计算出参数估计值 $\hat{A}$、$\hat{b}$，又由 $\hat{a} = e^{\hat{A}}$ 得到 $\hat{a}$。因此有回归方程

$$y = \hat{a} e^{bx}$$。

4 倒指数函数曲线 $y = ae^{b/x}$，其中 $a > 0$。

两边取常用对数：$\lg y = \lg a + \dfrac{b}{x}$，再作代换 $u = \dfrac{1}{x}$，$v = \lg y$，$A = \lg a$，则倒指数函数曲线方程就变成直线方程

$$v = A + bu$$。

由观测值 (x_i, y_i)，$i = 0,1,\cdots,n$ 按 $u_i = 1/x_i$，$v_i = \lg y_i$ 算出 (u_i, v_i)，$i = 0,1,\cdots,n$。对 u 与 v 利用本节前面的经验回归直线公式，计算出参数估计值 $\hat{A}$、$\hat{b}$，又由 $\hat{a} = e^{\hat{A}}$ 得到 $\hat{a}$。因此有回归方程

$$y = \hat{a} e^{b/x}$$。

5 对数函数曲线 $y = a + b\lg x$，其中 $x > 0$。

作代换 $u = \lg x$，$v = y$，则对数函数曲线方程就变成直线方程

$$v = a + bu$$。

由观测值 (x_i, y_i)，$i = 0,1,\cdots,n$ 按 $u_i = \lg x_i$，$v_i = y_i$ 算出 (u_i, v_i)，$i = 0,1,\cdots,n$。对 u 与 v 利用本节前面的经验回归直线公式，计算出参数估计值 $\hat{a}$、$\hat{b}$。由此得到回归方程

$$y = \hat{a} + \hat{b} \lg x$$。

6 S型函数曲线 $y = \dfrac{1}{a + be^{-x}}$。

上述方程等价于 $\dfrac{1}{y} = a + be^{-x}$，作代换 $u = e^{-x}$，$v = \dfrac{1}{y}$，则S型函数曲线方程就变成直线方程

$$v = a + bu$$。

由观测值 (x_i, y_i)，$i = 0,1,\cdots,n$ 算出 (u_i, v_i)，$i = 0,1,\cdots,n$。对 u 与 v 利用本节前面的经验回归直线公式，计算出参数估计值 $\hat{a}$、$\hat{b}$。因此有回归方程

$$y = \frac{1}{\hat{a} + \hat{b} e^{-x}}$$。

注：对于非线性回归问题的Matlab实现问题，一种方法是化为相应的线性模型实现，另一

种方法是直接应用Matlab中相应的命令，其结果是一致的。详见本节第五部分。

三　多元线性回归分析

一般地，在实际问题中影响应变量 y 的自变量往往不止一个，不妨设有 k 个为 x_1， x_2，…， x_k。通过观测得到一组（$k+1$ 维）相互独立的试验观测数据 $(x_{1j},x_{2j},\cdots,x_{kj},y_j)$，$j=1,2,\cdots,n$，其中 $n>k+1$。假设变量 y 与变量 x_1， x_2，…， x_k 之间有线性关系：

$$y=\beta_0+\beta_1x_1+\cdots+\beta_kx_k+\varepsilon, \tag{10.13}$$

其中 ε 是随机变量，一般假设 $E\varepsilon=0$， $D\varepsilon=\sigma^2>0$。则观测数据满足

$$y_j=\beta_0+\beta_1x_{1j}+\cdots+\beta_kx_{kj}+\varepsilon_j,\quad j=1,2,\cdots,n, \tag{10.14}$$

其中 ε_1，…， ε_n 互不相关且均是与 ε 同分布的随机变量。令

$$Y=\begin{bmatrix}y_1\\y_2\\\vdots\\y_n\end{bmatrix},\quad X=\begin{bmatrix}1&x_{11}&\cdots&x_{k1}\\1&x_{12}&\cdots&x_{k2}\\\vdots&\vdots&\vdots&\vdots\\1&x_{1n}&\cdots&x_{kn}\end{bmatrix},\quad \beta=\begin{bmatrix}\beta_0\\\beta_1\\\vdots\\\beta_k\end{bmatrix},\quad \varepsilon=\begin{bmatrix}\varepsilon_1\\\varepsilon_2\\\vdots\\\varepsilon_n\end{bmatrix},$$

则(10.14)可简写为

$$Y=X\beta+\varepsilon, \tag{10.15}$$

其中 X 为已知的 $n\times(k+1)$ 矩阵，称为回归设计矩阵或资料矩阵，Y 是 n 维观察值列向量，β 为 $k+1$ 维未知的列向量，ε 是满足

$$\begin{cases}E\varepsilon=0_1,\\COV(\varepsilon,\varepsilon)=\sigma^2I_n\end{cases}$$

的 n 维随机列向量，其中 0_1 为 n 维零向量，σ^2 是未知参数，$D\varepsilon_j=\sigma^2$，$j=1,2,\cdots,n$，I_n 为 n 阶单位矩阵，即对随机误差 ε_1，…， ε_n 作无偏、等方差与互不相关的假定。一般称

$$\begin{cases}Y=X\beta+\varepsilon,\\E\varepsilon=0_1,\\COV(\varepsilon,\varepsilon)=\sigma^2I_n\end{cases}, \tag{10.16}$$

为 k 线性回归模型（高斯—马尔科夫线性模型），并简记为 $(Y,X\beta,\sigma^2I_n)$。

对(10.16)取数学期望得到

$$y=\beta_0+\beta_1x_1+\cdots+\beta_kx_k,$$

称为线性回归方程。

对线性模型 $(Y,X\beta,\sigma^2I_n)$，所要考虑的主要问题是：

（i）用实验观测数据对未知参数 β_1，…， β_k 和 σ^2 做点估计和假设检验，从而建立因变量 y 和自变量 x_1，…， x_k 之间的线性关系；

（ii）在 $x_1=x_{10}$，…， $x_k=x_{k0}$ 处对 y 的值作预测和控制，并对 y 作区间估计。本部分总是假设 $n>k+1$。

1 未知参数β_i和σ^2估计

首先用最小二乘法求β_1，…，β_k的估计量。做误差平方和

$$Q=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_{1i}-\cdots-\beta_kx_{ki})^2,$$

寻求使得Q达到最小值的$\ddot{\beta}_1$，…，β_k作为β_1，…，β_k的点估计。由微分学中求最值的方法知，只需求解法方程组（或称正规方程组）

$$X^TX\ddot{\beta}=X^TY,$$

其中X、Y同前，$\beta=(\beta_0,\beta_1,\cdots,\beta_k)'$。解的估计值$\ddot{\beta}=(X^TX)^{-1}X^TY$，这样得到的$\beta_i$带入线性回归方程，得

$$\hat{y}=\beta_0+\beta_1x_1+\cdots+\beta_kx_k,$$

称为经验线性回归方程，β_i成为经验回归系数。同时有结论：

（i）β_i服从正态分布，$\ddot{\beta}$服从$k+1$维正态分布；

（ii）$\ddot{\beta}$是β的无偏估计且是最优线性无偏估计量，协方差矩阵为$\sigma^2(X^TX)^{-1}$。

（iii）$Q=(Y-G\ddot{\beta})^T(Y-G\ddot{\beta})$，其中$G=[1,x_1,\cdots,x_k]$。则有$E(Q)=(n-k-1)\sigma^2$，且

$$\frac{Q}{\sigma^2}\sim\chi^2(n-k-1)。$$

从而$E(\frac{Q}{n-k-1})=\sigma^2$，即$\frac{Q}{n-k-1}=\hat{\sigma}^2$为$\sigma^2$的无偏估计量。$\hat{\sigma}^2$是剩余方差，$\hat{\sigma}$是剩余标准差。

（iv）对Y的样本方差$S=\sum_{i=1}^{n}(y_i-\bar{y})^2$，有

$$S=Q+U,\quad U=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2,$$

其中Q是残差平方和，反映随机误差对y的影响，U是由回归方程引起的，称为回归平方和，反映自变量对y的影响。

2 多元线性回归中的假设检验

在实际问题中，往往事先不知道或不能确定随机变量y与自变量x_1，…，x_k之间确有线性关系。因而(10.13)往往是一种假设，因此在求出线性回归方程之后，还必须对求出的线性回归方程同实际观测数据拟合效果进行检验。类似于一元线性回归，可提出以下原假设

$$H_0:\beta_0=\beta_1=\cdots=\beta_k=0。$$

当H_0成立时，若令$R=\sqrt{\frac{U}{U+Q}}$（称R为相关系数），则有

$$F=\frac{U/k}{Q/(n-k-1)}=\frac{n-k-1}{k}\frac{R^2}{1-R^2}\sim F(k,n-k-1)。$$

当显著水平α给定后，由$\alpha = P\{F > F_{1-\alpha}(k,n-k-1) \mid H_0\text{成立}\}$，即

$$1-\alpha = P\{F \le F_{1-\alpha}(k,n-k-1) \mid H_0\text{成立}\}$$

知F检验法的检验规则为：

如果$F > F_{1-\alpha}(k,n-k-1)$，则拒绝$H_0$，认为因变量$y$与自变量$x_1$，…，$x_k$之间的线性关系显著；否则，认为$y$与$x_1$，…，$x_k$之间的线性关系不显著。

需要注意的是，y与x_1，…，x_k之间的线性关系不显著，可能出现几种情况：如y于其中某些自变量无关系，可以去掉这些自变量；y与x_1，…，x_k之间的存在非线性关系；还有其它变量与y有关系等。当然还有其它检验方法。

3 多元线性回归中的预测

（i）点预测

当我们求出回归方程

$$\hat{y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

并经过检验后，对于给定的自变量的值$x_1 = x_{10}$，…，$x_k = x_{k0}$，自然用$\hat{y}_0 = \beta_0 + \beta_1 x_{10} + \cdots + \hat{\beta}_k x_{k0}$来预测$y_0 = \beta_0 + \beta_1 x_{10} + \cdots + \beta_k x_{k0} + \varepsilon$。称$\hat{y}_0$为$y_0$的点预测。实际上，$\hat{y}_0$是$Ey_0$的无偏估计量（因为$E(\hat{y}_0 - y_0) = 0$）。

（ii）区间预测

由于$\hat{\beta} \sim N_{k+1}(\beta, \sigma^2 (X^T X)^{-1})$，所以$\hat{y} \sim N(G\beta, G\sigma^2 (X^T X)^{-1} G^T)$。当$G$已知时，由于$Q$与$\hat{\beta}$相互独立，从而$Q$与$\hat{y} = G\hat{\beta}$相互独立。又因为$y_1$，…，$y_n$相互独立，所以$y$与$\hat{y}$、$y$与$Q$相互独立。从而$y - \hat{y}$与$Q$相互独立，且有

$$y - \hat{y} \sim N(0, \sigma^2 (1 + G(X^T X)^{-1} G^T))。$$

令$|y - \hat{y}| \le K$，即y的置信区间为$(\hat{y} - K, \hat{y} + K)$，$K$由置信度$1-\alpha$确定，则当显著水平$\alpha$给定后，根据$\dfrac{Q}{\sigma^2} \sim \chi^2(n-k-1)$有

$$T = \frac{y - \hat{y}}{\sigma\sqrt{1 + G(X^T X)^{-1} G^T}} \Bigg/ \hat{\sigma} \sim t(n-k-1)。$$

从而

$$K = \hat{\sigma} t_{1-\alpha/2}(n-k-1)\sqrt{1 + G(X^T X)^{-1} G^T}。$$

因此当显著水平α给定后，y的置信区间为$(\hat{y} - K, \hat{y} + K)$，其中$K$由上式确定。

注：多元线性回归的Matlab实现详见本节第五部分。

四　逐步线性回归分析

从多元线性回归分析中我们知道，采用的自变量越多，则回归平方和越大，残差平方和越小。然而，采用较多的变量来拟合回归方程，得到的方程稳定性差，每个自变量的区间误差的积累将影响总体误差，用这样建立起来的回归方程作预测的可靠性差、精度低。另一方面，如果采用了对因变量影响小的自变量而遗漏了重要变量，可导致估计量产生偏倚和不一致性。因而

希望得到最优的回归方程。

逐步线性回归分析方法就是一种自动从大量可供选择的变量中选择那些对建立回归方程比较重要的变量的方法，它是在多元线性回归基础上派生的一种算法技巧，详可参阅相应的文献。其基本思路为：从一个自变量开始，视自变量对y作用的显著程度，从大到小依次逐个引入回归方程。当引入的自变量由于后面自变量的引入而变得不显著时，要将其剔除掉。引入一个自变量或从回归方程中剔除一个自变量，为逐步回归的一步。对于每一步，都要进行y值检验，以确保每次引入新的显著性变量前回归方程中只包含对y作用显著的变量。这个过程反复进行，直至即无不显著的变量从回归方程中剔除，又无显著变量可引入回归方程止。

使用逐步线性回归时要注意：要适当选择引入变量的显著性水平和剔除变量的显著性水平；应尽量选择那些相互独立性强的变量。

五 回归分析的Matlab实现

Matlab统计工具箱中提供了一些回归分析的命令，现介绍如下。

1 多元线性回归

多元线性回归的命令是regress，此命令也可用于一元线性回归。其格式为：

（1）确定回归系数的点估计，用命令：b=regress(Y，X)。

（2）求回归系数的点估计和区间估计，并检验回归模型，用命令：

[b，bint，r，rint，stats]=regress(Y，X，alpha)。

（3）画出残差及其置信区间，用命令：

rcoplot（r，rint）。

在上述命令中，各符号的含义为：

（i）$b=\hat{\beta}$，Y，X的定义同本部分前面所述。对一元线性回归，在$b=\hat{\beta}$，Y，X中取k=1即可；

（ii）alpha为显著性水平（缺省时为0.05）；

（iii）bint为回归系数的区间估计；

（iv）r与rint分别为残差及其置信区间；

（v）stats是用于检验回归模型的统计量，有三个数值，第一个是R^2，第二个是F值，第三个是与F对应的概率P。其中R^2与F定义同前，值越大，说明回归方程越显著，$P<\alpha$时拒绝H_0，回归模型成立。

例7 合金的强度y与其中的碳含量x有比较密切的关系，今从生产中收集了一批数据如下表。试先拟合一个函数$y(x)$，再用回归分析对它进行检验。

x	0.10	0.11	0.12	0.13	0.14	0.15	0.16	0.17	0.18
y	42.0	41.5	45.0	45.5	45.0	47.5	49.0	55.0	50.0

解 先画出散点图：

```
x=0.10:0.01:0.18;
```

```
y=[42.0,41.5,45.0,45.5,45.0,47.5,49.0,55.0,50.0];
polt(x,y,'+')
```

可知y与x大致为线性关系。

设回归模型为 $y=\beta_0+\beta_1 x$ ，用regress和rcoplot编程如下：

```
clc,clear
x1=[0.10:0.01:0.18]';
y=[42.0,41.5,45.0,45.5,45.0,47.5,49.0,55.0,50.0]';
x=[ones(9,1),x1];
[b,bint,r,rint,stats]=regress(y,x);
b,bint,stats,rcopolt(r,rint)
```

得到

```
b=27.4722    137.5000
bint=18.6851      36.2594
       75.7755     199.2245
stats=0.7985     27.7469    0.0012
```

即 $\hat{\beta}_0$ =27.4722， $\hat{\beta}_1$=137.5000， $\hat{\beta}_0$ 的置信区间是[18.6851，36.2594]， $\hat{\beta}_1$ 的置信区间是[75.7755，199.2245]； R^2=0.7985， F =27.7469， $p=$0.0012。

可知所设回归模型成立。

观察命令rcoplot(r,rint)所画的残差分布，除第8个数据外其余残差的置信区间均包含零点，第8个点应视为异常点，将其剔除后重新计算，可得

```
b=30.7280    109.3985
bint=26.2805      35.2834
       76.9014     141.8955
stats=0.9188     67.8534    0.0002
```

应该用修改后的这个结果。

例8 某厂生产的一种电器的销售量 y 与竞争对手的价格x1和本厂的价格x2有关。下表是该厂商品在10个城市的销售记录。试根据这些数据建立y与x1和x2的关系式，对得到的模型和系数进行检验。若某市本厂产品售价160（元），竞争对手售价170（元），预测商品在该市的销售量。

x1元	120	140	190	130	155	175	125	145	180	150
x2元	100	110	90	150	210	150	250	270	300	250
Y个	102	100	120	77	46	93	26	69	65	85

解 分别画出y关于x1和y关于x2的散点图，可以看出y与x2有较明显的线性关系，而y与x1之间的关系则难以确定，我们将作几种尝试，用统计分析决定优劣。

设回归模型为

$$y=\beta_0+\beta_1 x_1+\beta_2 x_2$$ 。

编写如下程序：

```
x1=[120  140  190  130  155  175  125  145  180  150]';
x2=[100  110   90  150  210  150  250  270  300  250]';
y=[102  100  120   77   46   93   26   69   65   85]';
x=[ones(10,1),x1,x2];
[b,bint,r,rint,stats]=regress(y,x);
b,bint,stats
```

得到

```
b=66.5176    0.4139    -0.2698
bint=-32.5060     165.5411
       -0.2018     1.0296
       -0.4611     -0.0785
stats=0.6527     6.5786    0.0247
```

可以看出结果不是太好，p=0.0247，取α=0.05时所设回归模型可用，但取α=0.01时所设回归模型不能用；R^2=0.6527较小；$\hat{\beta}_0$，$\hat{\beta}_1$的置信区间包含了零点。下面将试图用x1，x2的二次函数改进它。

2 多项式回归

（1）一元多项式的回归和预测

一元多项式的回归和预测可用命令polyfit或polytool和polyval或polyconf来实现。其命令格式如下：

令$x=(x_1,x_2,\cdots,x_n)$，$y=(y_1,y_2,\cdots,y_n)$，$P=(a_1,a_2,\cdots,a_{m+1})$是多项式$y=a_1x^m+a_2x^{m-1}+\cdots+a_mx+a_{m+1}$的系数，$S$是一个矩阵，用来估计预测误差。

回归可以用命令[P,S]=polyfit(x,y,m)或polytool(x,y,m)实现，其中[P,S]=polyfit(x,y,m)是确定多项式系数的命令；命令polytool(x,y,m)命令产生一个交互式的画面，在画面中绿色曲线为拟合曲线，它两侧的红线是y的置信区间。可以用鼠标移动图中的十字线来改变图下方的x值，也可以在窗口内输入，左边就给出y的预测值与置信区间。通过左下方的Export下拉式菜单，可以输出回归系数等。

预测和预测误差估计的命令为polyval或者polyconf，其中Y=polyval(p,x)求polyfit所得的回归多项式在x处的预测值Y；[Y,DELTA]=polyconf(p,x,S,alpha)求polyfit所得的回归多项式在x处的预测值Y及预测值的显著性为1-alpha的置信区间Y$\pm$DELTA；alpha缺省时为0.05。

一元多项式回归也可化为多元线性回归来解。

例9 观测物体降落的距离 y 与时间 x 的关系，得到数据如下表，求 y 关于 x 的回归方程 $y = a + bx + cx^2$ 。

x	1/30	2/30	3/30	4/30	5/30	6/30	7/30
y	11.86	15.67	20.60	26.67	33.71	41.93	51.13
x	8/30	9/30	10/30	11/30	12/30	13/30	14/30
y	61.49	72.90	85.44	99.08	113.77	129.54	146.48

解　方法（一）　用一元多项式回归，编写程序如下：

```
x=1/30:1/30:14/30;
y=[11.86  15.67  20.60  26.67  33.71  41.93  51.1361.49  72.90  85.44   99.08
    113.77  129.54  146.48];
    [p,S]=polyfit(x,y,2);
```

得到p=

489.2946　65.8896　9.1329

即a=9.1329　b=65.8896　c=489.2846。

方法（二）　化为多元线性回归，其程序为：

```
x=1/30:1/30:14/30;
y=[11.86  15.67  20.60  26.67  33.71  41.93  51.1361.49  72.90  85.44   99.08
113.77  129.54  146.48];
T=[ones(14,1),t,t.^2]
[b,bint,r,rint,stats]=regress(y,T);
b,stats
```

得到结果：b=9.1329　65.8896　489.2946

stats=1.0e+007*

0.0000　1.0378　0

可以看出，两种方法的出的结果是一样的。

（2）多元二项式回归

多元二项式回归可用命令：rstool(x,y,model,alpha)。其中，输入数据x、y分别为 $n \times m$ 矩阵和n维列向量；alpha为显著性水平（缺省时为0.05）；model由下列4个模型中选择1个（用字符串输入，缺省时为线性模型）：

linear（线性）：$y = \beta_0 + \beta_1 x + \cdots + \beta_m x^m$；

purequadratic（纯二次）：$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \sum_{j=1}^{m} \beta_{jj} x_j^2$；

interaction（交叉）：$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \sum_{1 \le j \ne k \le m} \beta_{jk} x_j x_k$；

quadratic（完全二次）： $y=\beta_0+\beta_1x_1+\cdots+\beta_mx_m+\sum_{1\le j,k\le m}\beta_{jk}x_jx_k$。

命令rstool产生一个交互式画面，画面中有m个图形，这m个图形分别给出了一个独立变量xi（另m-1个变量取固定值）与y的拟合曲线，以及y的置信区间。可以通过键入不同的xi的值来获得相应的y值。

图的左下方有两个下拉式菜单，一个菜单Export用以向Matlab工作区传送数据，包括beta（回归系数）、rmse（剩余标准差）、residuals（残差）。另一个菜单model用以在上述4个模型中选择。可以分别选4个模型，并比较它们的剩余标准差，其中最接近于0的模型是最好的。

我们再作一遍例8商品销售量与价格问题，选择纯二次模型，即

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_1^2+\beta_4x_2^2$$ 。

编程如下：

```
x1=[120  140  190  130  155  175  125  145  180  150]';
x2=[100  110   90  150  210  150  250  270  300  250]';
y=[102  100  120   77   46   93   26   69   65   85]';
x=[x1 x2];
rstool(x,y,'purequadratic')
```

得到一个交互式画面，给出两幅图形。左边图形是x1固定时的曲线y(x1)及其置信区间，右边图形是x2固定时的曲线y(x2)及其置信区间。用鼠标移动图中的十字线，或在图下方窗口内输入，可改变x1、x2。画面左边给出y的预测值即其置信区间，用这种画面可以回答例8提出的“若谋市本厂产品售价160（元），竞争对手售价170（元），预测商品在该市的销售量”问题。

在画面左下方的下拉式菜单Export中选择“all”，则beta、rmse和residuals都传送到Matlab工作区中。在Matlab工作区中输入命令：

beta,rmse

得到结果：beta=-312.5871 7.2701 -1.7337 -0.0228 0.0037

rmse=16.6436

如果在另一菜单model选择其它多元二项式模型，比较它们的剩余标准差就会发现，本例的所选模型的rmse=16.6436最小。

注：本例中的模型亦可化为多元线性回归来做。请读者自己编程并比较结果。

3 非线性回归

非线性回归可用命令nlinfit,nlintool,nlparci,nlpredci来实现。命令格式如下：

回归：回归可用命令[beta,r,J]=nlinfit(x,y,model,beta0)或者nlintool(x,y,model,beta0,alpha)来实现。其中命令[beta,r,J]=nlinfit(x,y,model,beta0)的作用为确定回归系数；而命令nlintool(x,y,model,beta0,alpha)产生一个交互式的画面，画面中有拟合曲线和y的置信区间。通过左下方的Export下拉式菜单，可以输出回归系数等。

这里的输入数据x、y分别为$n \times m$矩阵和n维列向量，对一元非线性回归，x为n维列向量；model是事先用m-文件定义的非线性函数；beta0是回归系数的初值。Beta是估计出的回归系数，r（残差）、J（Jacobian矩阵）是估计预测误差需要的数据。alpha为显著性水平，缺省时为0.05。

预测和预测误差估计：预测和预测误差估计的命令格式为

betaci=nlparci(beta,r,J) 其用途为记算回归系数的置信区间；

[Y,DELTA]=nlpredci(model,x,beta,r,J) 其用途为求nlinfit或nlintool所得的回归函数在x处的预测值Y及预测值的显著性为1-alpha的置信区间Y$\pm$DELTA；alpha缺省时为0.05。

某些非线性回归也可化为多元线性回归来解。

例10 在研究化学动力学反应过程中，建立了一个反应速度和反应物含量的数学模型，形式为

$$y = \frac{\beta_4 x_2 - x_3 / \beta_5}{1 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3},$$

其中β_1，…，β_5式未知系数，x_1, x_2, x_3是三种反应物（氢，n戊烷，异构戊烷）的含量，y是反应速度。今测的一组数据如下表，试由此确定参数β_1，…，β_5，并给出置信区间。β_1，…，β_5的参考值为（0.1,0.05,0.02,1,2）。

序号	反应速度y	氢x1	n戊烷x2	异构戊烷x3
1	8.55	470	300	10
2	3.79	285	80	10
3	4.82	470	300	120
4	0.02	470	80	120
5	2.75	470	80	10
6	14.39	100	190	10
7	2.54	100	80	65
8	4.35	470	190	65
9	13.00	100	300	54
10	8.50	100	300	120
11	0.05	100	80	120
12	11.32	285	300	10
13	3.12	285	190	120

解　首先，以回归系数和自变量为输入变量，并将要拟合的模型写成函数文件huaxue.m:

```
function yhat=huaxue(beta,x);
yhat=(beta(4)*x(2)-x(3)/beta(5))./(1+beta(1)*x(1)+beta(2)*x(2)+beta(3)*x(3));
```

然后，用nlinfit计算回归系数，用nlparci计算回归系数的置信区间，用nlpredci计算预测值即其置信区间，编程如下：

```
clc,clear
x0=[1    8.55    470    300    10
    2    3.79    285    80     10
    3    4.82    470    300    120
    4    0.02    470    80     120
```

```
     5      2.75     470     80      10
     6      14.39    100     190     10
     7      2.54     100     80      65
     8      4.35     470     190     65
     9      13.00    100     300     54
     10     8.50     100     300     120
     11     0.05     100     80      120
     12     11.32    285     300     10
     13     3.12     285     190     120];
x=x0(:,3:5);
y=x0(:,2);
beta=[0.1,0.05,0.02,1,2]; %回归系数的初值
[betahat,f,j]=nlinfit(x,y,'huaxue',beta); % f,j是下面命令用的信息
betaci=nlparci(betahat,f,j);
betaa=[betaa,betaci]  %回归系数及其置信区间
[yhat,delta]=nlpredci('huaxue',x,betahat,f,j) %y的预测值及其置信区间半径，置信区间为yhat±delta。
```

用命令nlinfit(x,y,'huaxue',beta)可看到画面，并传出剩余标准差rmse=0.1933。

4 逐步回归

逐步回归的命令是stepwise，它提供了一个交互式画面，通过此工具可以自由地选择变量，进行统计分析。通常用法是：

stepwise(x，y，inmodel，alpha)，

其中x是自变量数据，y是因变量数据，分别为$n\times m$和$n\times 1$矩阵，inmodel是矩阵的列数指标，给出初始模型中包括的子集（缺省时设定为全部自变量），alpha为显著水平（缺省时为0.05）。

运行stepwise命令时产生三个图形窗口：Stepwise Plot，Stepwise Table，Stepwise History。所有这些图形界面都由热区，即当鼠标移到图形的某个区域时，鼠标的指针会变成一个小圆，点击后会产生交互作用。

在Stepwise Plot窗口，显示出各项的回归系数及其置信区间。其中：点表示回归系数的值，点两边的水平（实或虚）直线段表示其置信区间（虚线表示该变量的拟合与0无显著差异，实线表示有显著差异）；绿色的线表示当前在模型中的项，红色的线表示当前不在模型中的项。点击一条线会改变其状态，即在模型中的项（绿线）会被移去（变为红线），不在模型中的项（红线）会被加入（变为绿线）。次窗口中的Export下拉式菜单可以向Matlab工作区传送各种数据。次窗口中的Scale Inputs可对输入数据的每列进行正态化处理，使其标准差为1。

在Stepwise Table窗口中列出了一个统计表，包括回归系数及其置信区间，以及模型的统计量剩余标准差(RMSE)、相关系数(R-square)、F值、与F值对应的概率P。

例11 水泥凝固时放出的热量y与水泥中4种化学成分x1,x2,x3,x4有关，今测得一组数据如下，试用逐步回归来确定一个线性模型。

序号	x1	x2	x3	x4	y
1	7	26	6	60	78.5
2	1	29	15	52	74.3
3	11	56	8	20	104.3
4	11	31	8	47	87.6
5	7	52	6	33	95.9
6	11	55	9	22	109.2
7	3	71	17	6	102.7
8	1	31	22	44	72.5
9	2	54	18	22	93.1
10	21	47	4	26	115.9
11	1	40	23	34	83.8
12	11	66	9	12	113.3
13	10	68	8	12	109.4

解　编程如下：

```
clc,clear
x0=[1     7     26    6     60     78.5
    2     1     29    15    52     74.3
    3     11    56    8     20     104.3
    4     11    31    8     47     87.6
    5     7     52    6     33     95.9
    6     11    55    9     22     109.2
    7     3     71    17    6      102.7
    8     1     31    22    44     72.5
    9     2     54    18    22     93.1
    10    21    47    4     26     115.9
    11    1     40    23    34     83.8
    12    11    66    9     12     113.3
    13    10    68    8     12     109.4];
 x0=(:,2:5);
 y=x0(:,6);
 stepwise(x,y)
```

得到图Stepwise Plot和Stepwise Table表。图Stepwise Plot中四条直线都是虚线，说明模型的显著

性不好，从Stepwise Table表中可以看出变量x3和x4显著最差。移去这两个变量，图Stepwise Plot中点击直线3和直线4，这两条直线变为红色，同时直线1和直线2变为实线，说明移去变量x3和x4后模型具有显著性。从新的统计结果可以看出，虽然剩余标准差（RMSE）没有太大的变化，但是统计量F的值明显增大，因此新的回归模型更好一些。

对变量y和x1、x2作线性回归：

stepwise(x,y,[1,2])

得到结果：y=52.5773+1.4683x_1+0.6623x_2。

§6 建模案例：水塔水流量的估计

一　问题的提出

某居民区供水机构有一供居民用水的圆柱形水塔，由于该机构没有测量流入或流出该水塔的水量的装置，水塔的管理者只能通过隔一段时间测量一次水塔的水位来估计水的流量。但面临的困难是，当水塔水位下降到设定的最低水位时，水泵自动启动向水塔供水，到设定的最高水位时停止供水，这段时间无法测量水塔的水位和水泵的供水量。通常每天水泵供水一到二次，每次约2个小时。

水塔是一个高为12.2米，直径为17.4米的正圆柱。按照设计，水塔水位降至约8.2米时，水泵自动启动，水位升到约10.8米是水泵停止工作。

表1是某一天管理人员测量该水塔的水位测量记录（符号“//”表示水泵启动）。请您帮助该管理人员估计一天中任何时刻（包括水泵正在供水时刻）从水塔中流出的水流量及一天的居民用水总量。

表1 水位测量记录

时刻(h)	0	0.921	1.843	2.949	3.871	4.978	5.900	7.006	7.928	8.967
水位(m)	9.677	9.479	9.308	9.125	8.982	8.814	8.686	8.525	8.388	8.220
时刻(h)	9.9811	10.925	10.954	12.032	12.954	13.875	14.982	15.903	16.826	17.931
水位(m)	//	//	10.820	10.500	10.210	9.936	9.653	9.409	9.180	8.921
时刻(h)	19.037	19.959	20.839	22.015	22.958	23.880	24.986	25.908		
水位(m)	8.662	8.433	8.220	//	10.820	10.591	10.354	10.180		

二　问题的分析

流量是单位时间流出的水的体积。由于水塔是正圆柱形，截面面积是常数，在水泵停止供水的时段，流量很容易从水位对时间的变换率算出，问题是如何估计水泵供水时段的水流量。

水泵供水时段的流量与供水时段前后的流量有关。如果能够通过测量数据，产生若干个时刻的水流量，则计算水流量问题就转化为插值问题。这样水泵供水时段的流量就可以用供水时段前后的流量插值得到。作为用于函数插值的原始数据我们希望水泵不工作时段的流量越准确越好。

有了任何时刻的流量，就不难计算一天的总用水量。水泵不工作时段的用水量可以由测量

记录直接得到，如由表1可知从t=0到t=8.967(h)水位下降了9.677-8.220=1.457(m)，乘以水塔的截面积就是这一时段的用水量。这个数值可以用来检验插值的结果。

注：此处我们只考虑水流量的插值问题，当然亦可以考虑水位的插值问题；或者亦可以将此问题按拟合问题进行处理。

三　模型假设

1 流量只取决于水位差，与水位本身无关。按照Torricelli定律从小孔流出的流体的流速（流量）正比于水面高度的平方根，问题给出水塔的最低和最高水位分别约是8.22米和10.82米（设出水口的水位为零），因为$\sqrt{10.82/8.22} \approx 1$，所以可以忽略水位对流速（流量）的影响。

2 水塔中水流量是时间的连续光滑函数，与水泵是否工作无关。

3 水泵是否工作完全取决于水塔内的水位的高度，且每次加水的工作时间约为2小时，从表1中可知两次供水时间段分别为[8.967，10.954]和[20.839，22.958]。

4 水泵工作时单位时间的供水量大致是常数，此常数大于单位时间的平均流量。

5 一天之中可以从任何时刻开始，即从t=0(h)到t=24(h)结束与从t=α(h)到t=24+α(h)结束差别不大。每天同一时刻水流量保持相同，并且一天中开始时刻的不同对一天中的总用水量影响很小。

6 水塔内部底面直径为17.4米。

四　数据处理

1 水体积计算

水塔是一个正圆柱，塔内水的体积为 $V = \frac{\pi}{4} D^2 h$，其中D为水塔直径，h为水面高度。近似地取π=3.141592654，得到不同时刻水塔中水的体积如表2。

表2 水塔中水的体积（单位：时刻（小时），体积（立方米））

时刻	0	0.921	1.843	2.949	3.871	4.978	5.900	7.006	7.928	8.967
水位	2294	2247	2206	2163	2129	2089	2059	2020	1988	1948
时刻	9.9811	10.925	10.954	12.032	12.954	13.875	14.982	15.903	16.826	17.931
水位	//	//	2564	2489	2420	2355	2288	2230	2176	2114
时刻	19.037	19.959	20.839	22.015	22.958	23.880	24.986	25.908		
水位	2053	1999	1948	//	2564	2510	2454	2413		

2 水流速(流量)的近似

水流速度（流量）是水塔中水的体积对时间的导数。由于没有水的体积关于时间的函数表达式，而只是一个离散的函数值表2，因此考虑常用的差商代替导数处理方法。为提高精度，采用二阶差商公式。具体地，表2的数据点被水泵两次工作分割成三组数据，对每组数据中间数据采用二阶中心差商，两边数据采用二阶向前或向后差商得到数据表3。其公式分别为：

中心差商公式 $$\nabla^2 v_i = \frac{-v_{i+2} + 8v_{i+1} - 8v_{i-1} + v_{i-2}}{12(t_{i+1} - t_i)}$$；

向前差商公式 $\nabla^2 v_i = \dfrac{-v_{i+2} + 4v_{i+1} - 3v_i}{2(t_{i+1} - t_i)}$；

向后差商公式 $\nabla^2 v_i = \dfrac{3v_i - 4v_{i-1} + v_{i-2}}{2(t_{i+1} - t_i)}$。

表3 水塔中水的流速（单位：时刻（小时），流速（立方米/小时））

时刻	0	0.921	1.843	2.949	3.871	4.978	5.900	7.006	7.928	8.967
流速	54.516	42.320	38.085	41.679	33.297	37.817	30.748	38.455	32.122	41.718
时刻	9.9811	10.925	10.954	12.032	12.954	13.875	14.982	15.903	16.826	17.931
流速	//	//	73.686	76.434	71.686	60.190	68.333	59.217	52.011	56.626
时刻	19.037	19.959	20.839	22.015	22.958	23.880	24.986	25.908		
流速	63.023	54.859	55.439	//	57.602	57.776	51.891	36.464		

五 模型建立与求解

1 做出水流速的散点图

Matlab程序如下：

```
t=[0  0.921  1.843  2.949  3.871  4.978  5.900  7.006  7.928  8.967 10.954 12.032 12.954 13.875 14.982
   15.903 16.826 17.931 19.037 19.959 20.839 22.958 23.880 24.986 25.908];
r=[54.516 42.320 38.085 41.679 33.297 37.817 30.748 38.455 32.122 41.718 73.686 76.434 71.686
   60.190 68.333 59.217 52.011 56.626 63.023 54.859 55.439 57.602 57.776 51.891 36.464];
plot(t,r,'+');
title('流速散点图'); xlabel('时间（小时）'); ylabel('流速（立方米/小时）')
```

做出的散点图略。

2 模型及计算结果

现在问题已转变为根据流速 $f(t)$ 的一个函数值表3，产生函数 $f(t)$ 在整个区间（二十四小时）上的函数或函数值，插值和拟合是最常用的两种方法。如果建立拟合模型，需要根据散点图的趋势，选择适当的拟合函数形式。若采用插值模型，可以考虑分段线性插值（光滑性较差）、分段三次多项式插值、三次样条插值等插值方法。我们采用插值方法。

通过对不同插值方法的比较，结合假设2，即流速似时间的连续光滑函数，下面采用三次样条插值模型。

1）水流量及一天用水总量　首先由三次样条插值计算得到水流量函数 $f(t)$。其Matlab程序为：

```
t=[0  0.921  1.843  2.949  3.871  4.978  5.900  7.006  7.928  8.967 10.954 12.032 12.954 13.875 14.982
   15.903 16.826 17.931 19.037 19.959 20.839 22.958 23.880 24.986 25.908];
r=[54.516 42.320 38.085 41.679 33.297 37.817 30.748 38.455 32.122 41.718 73.686 76.434 71.686
   60.190 68.333 59.217 52.011 56.626 63.023 54.859 55.439 57.602 57.776 51.891 36.464];
[1,n]=size(t); d1=t(n)-t(1);
x=t(0):1/3600:t(n);  %被插值点
ys=interp1(t,r,x,'spline'); %样条插值输出
plot(x,ys);
```

title('样条插值下的水流速图'); xlabel('时间（小时）'); ylabel('流速（立方米/小时）')

做出的水流速曲线图略。

2）一天（24小时）的用水总量　用三次样条插值模型得到的函数$f(t)$在区间[0,24]上积分得到的结果为1257.3立方米。

3）误差分析　直接由表3的水流速数据，用梯形公式进行数值积分得到的结果为1250.3立方米，与上面得到的结果比较，则有：三次样条插值的绝对误差为7个立方米，相对误差为0.56%（不到百分之一）。

另外，测量数据的误差可控制在0.5%；数值微分、数值积分的误差亦可以控制在一定的范围内。

六　模型的稳定性分析与检验

1 稳定性分析

用不同时刻作为起始点，使用三次样条插值函数得到的水流量函数$f(t)$，在长度为24小时的时间区间上积分，所得结果相差无几，详见表4。这说明我们所建立的数值微分模型和三次样条插值模型是稳定的。

表4 不同起点计算出的24小时用水量（单位：立方米）

起始点	0.00	0.25	0.50	0.75	1.00	1.25	1.50	1.75
用水量	1257.3	1258.5	1260.3	1262.6	1265.1	1267.6	1269.6	1270.7

2 检验

1）　分三段（水泵未工作）的实际用水量与由模型推算的用水量的差异很小，见表5，最大的不超过4%，三段总计不超过0.5%。其中由模型推算的用水量为水流量函数$f(t)$的积分；实际用水量为水塔内水的体积之差。

表5 分三段的实际用水量与模型计算出的24小时用水量比较（单位：立方米）

	实际用水量	模型用水量	绝对误差	相对误差
第一段[0，8.967]	345.3682	338.0511	7.3171	2.1%
第二段[10.954，20.839]	616.3161	620.8341	4.5180	0.7%
第三段[22.958，25.908]	151.7308	156.6000	4.8692	3.2%
三段总计	1113.4151	1115.4852	2.0701	0.2%

2）两次充水期间，水泵注入量的差异大约为3个立方米，不到0.5%。

水泵充水量=充水后的水量+充水期间的流出量—充水前的水量。

第一次水泵充水量=2564+117.4463-1948=733.4463立方米；

第二次水泵充水量=2564+120.7052-1948=736.7052立方米。

由于水泵功率=充水量/充水时间，由上述计算知，水泵的功率大约为367立方米/小时，而且两次冲水期间计算出的功率大致相等，这与实际问题是一致的。

3）用水高峰的比较

实际与模型之间相差无几。实际用水高峰可近似地用差商最大值点表示为t=11，即上午11点钟左右；模型得到的用水高峰可由模型得到的水流量函数依次在单位区间上积分或者找出水流量函数的最大值点得出，它们都在11点左右。

七　模型的优缺点及推广

1 优点

（1）模型灵活性好、稳定性强，可用于那些拥有地方性的竖直圆柱型水塔的小城镇和乡镇。模型中的输入数据可以是任何近似均匀的时间间隔时的水位，时间间隔大约2小时。

（2）模型中的数学概念简单，并且容易理解。只用到数值计算知识。

（3）模型容易实现，且给出了一天里水流速度和用水量的精确估计。

2 缺点

（1）本模型受水塔的几何形状限制。

（2）光滑曲线的逼近方法不能模拟真实流动中流速的微小变化，实际流动中流速可能会有一种高程度的“噪音”，即激波出现。

3 改进与推广

（1）我们可以在模型中用一个参数来限定不同几何形状的水塔。

（2）可以通过对流速数据进行回归分析等一系列处理，以便得到一些随机变化的特征。

习　题　十

1. 用给定的多项式，如 $y=x^3-6x^2+5x-3$，产生一组数据 (x_i, y_i)，$i=1,2,\cdots,m$，再在 y_i 上添加随机干扰（可用rand产生（0,1）均匀分布随机数，或用randn产生 $N(0,1)$ 分布随机数），然后用 x_i 和添加了随机干扰的 y_i 作3次多项式拟合，与原系数比较。如果作2或4次多项式拟合，结果如何？

2. 已知平面区域 $0\le x\le 5600$，$0\le y\le 4800$ 的高程数据见下表（单位：m）。

4800	1350	1370	1390	1400	1410	960	940	880	800	690	570	430	290	210	150
4400	1370	1390	1410	1430	1440	1140	1110	1050	950	820	690	540	380	300	210
4000	1380	1410	1430	1450	1470	1320	1280	1200	1080	940	780	620	460	370	350
3600	1420	1430	1450	1480	1500	1550	1510	1430	1300	1200	980	850	750	550	500
3200	1430	1450	1460	1500	1550	1600	1550	1600	1600	1600	1550	1500	1500	1550	155
2800	0														
2400	950	1190	1370	1500	1200	1100	1550	1600	1550	1380	1070	900	1050	1150	120
2000	0														
1600	910	1090	1270	1500	1200	1100	1350	1450	1200	1150	1010	880	1000	1050	110
1200	0														
800	880	1060	1230	1390	1500	1500	1400	900	1100	1060	950	870	900	936	950
400	830	980	1180	1320	1450	1420	400	1300	700	900	850	810	380	780	750
0	740	880	1080	1130	1250	1280	1230	1040	900	500	700	780	750	650	550
	650	760	880	970	1020	1050	1020	830	800	700	300	500	550	480	350
	510	620	730	800	850	870	850	780	720	650	500	200	300	350	320
	370	470	550	600	670	690	670	620	580	450	400	300	100	150	250
Y/X	0	400	800	1200	1600	2000	2400	2800	3200	3600	4000	4400	4800	5200	5600

试用二维插值求 x, y 方向间隔都为50的高程，并画出该区域的等高线。

3. 用最小二乘法求一形如 $y = ae^{bx}$ 的经验公式拟合下列数据

x_i	1	2	3	4	5	6	7	8
y_i	15.3	20.5	27.4	36.6	49.1	65.6	87.87	117.6

4. 某人记录了21天使用空调器的时间和使用烘干器的次数，并监视电表以计算出每天的耗电量，数据见下表，试研究耗电量（KWH）与空调器使用的小时数（AC）和烘干器使用次数（DRYER）之间的关系，建立并检验回归模型，诊断是否有异常点。

序号	1	2	3	4	5	6	7	8	9	10	11
KWH	35	63	66	17	94	79	93	66	94	82	78
AC	1.5	4.5	5.0	2.0	8.5	6.0	13.5	8.0	12.5	7.5	6.5
DRYER	1	2	2	0	3	3	1	1	1	2	3

序号	12	13	14	15	16	17	18	19	20	21
KWH	65	77	75	62	85	43	57	33	65	33
AC	8.0	7.5	8.0	7.5	12.0	6.0	2.5	5.0	7.5	6.0
DRYER	1	2	2	1	1	0	3	0	1	0

5. 在一丘陵地带测量高程，x和y方向每隔100米测一个点，得高程如下表，试拟合一曲面，确定合适的模型，并由此找出最高点和该点的高程，给出回归分析。

y \ x	100	200	300	400	500
100	636	697	624	478	450
200	698	712	630	478	420
300	680	674	598	412	400
400	662	626	552	334	310

6. 一矿脉有13个相邻样本点，人为地设定一原点，现测的各样本点对原点的距离x，与该样本点处某种金属含量y的一组数据如下，画出散点图观测二者之间的关系，试建立合适的回归模型，如二次曲线、双曲线、对数曲线等。

x	2	3	4	5	7	8	10
y	106.42	109.20	109.58	109.50	110.00	109.93	110.49

x	11	14	15	16	18	19	21
y	110.59	110.60	110.90	110.76	110.00	111.20	110.53

参 考 文 献

[1] 赵静，但琦主编，严尚安，杨秀文副主编，数学建模与数学实验，北京：高等教育出版社；海德堡：施普林格出版社，2000。

[2] 张德荣，王新民，高安民，计算方法与算法语言，北京：高等教育出版社，1981。

[3] 易大义，陈道琦，数值分析引论，杭州：浙江大学出版社，1998。

[4] 重庆大学数学系组编，傅鹂，龚劬，刘琼荪，何中市编著，数学实验，北京：科学出版社，2000。

[5] 司守奎主编，徐珂文，李日华副主编，数学建模（海军航空工程学院使用教材），2001。

[6] 中山大学数学力学系《概率论及数理统计》编写小组编，概率论及数理统计（上），北京：高等教育出版社，1980。